DOCUMENT RESUME

ED 173 438 TM 009 577

TITLE	Proceedings of the Invitational Conference on Testing Problems (24th, New York, New York, November 2, 1963).
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	2 Nov 63
NOTE	161p.
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Academic Achievement; *Cognitive Ability; College Students; *Educational Testing; Equated Scores; Higher Education; Medical Education; Norm Referenced Tests; *Norms; Personality Assessment; *Predictive Measurement; Social Responsibility; Student Testing; Success Factors; *Testing Problems; *Test Reliability; Test Validity

ABSTRACT

Conference speakers reviewed and analyzed the current thinking underlying the basic concepts of test norms, reliability, and validity. Roger T. Lennon's paper called for more attention to the development of norming theory and summarized current norming practices and the establishment of a system which would permit comparable norms for tests which were standardized on different samples. Robert L. Thorndike commented on proposals, practices, and procedures for estimating reliability. Anne Anastasi reviewed recent approaches to the study of validity. During the session devoted to test use in medicine, John P. Hubbard described a method used to test the diagnostic competence of interns. E. Lowell Kelly reported on an extensive investigation of predictor and criterion variables of concern to medical schools. Jerome S. Bruner presented the luncheon address on human development. Warren G. Findley discussed current theory and applications in the appraisal of cognitive ability. Non-cognitive aspects of student performance were treated by Samuel Messick, who considered the portential contribution of personality assessment techniques to the prediction of college success. The final speaker was Robert L. Ebel, who spoke on the social consequences of educational testing. (BH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED173438

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PROCEEDINGS

# of the
1963
Invitational
Conference
on
Testing
Problems

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Educational
Testing Service
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM009 577

Copyright © 1964 by Educational Testing Service. All rights reserved.

Library of Congress Catalogue Number: 47-11220
Printed in the United States of America

Invitational Conference on Testing Problems

November 2, 1963
Hotel Roosevelt
New York City

ALEXANDER G. WESMAN
Chairman

EDUCATIONAL TESTING SERVICE
Princeton, New Jersey
Berkeley, California

ETS
Board of Trustees
1963-64

Katharine E. McBride, *Chairman*

John S. Allen
George F. Baughman
Frank H. Bowles
Samuel M. Brownell
John H. Fischer
John W. Gardner
Calvin E. Gross
Mason W. Gross
David D. Henry
Arnold E. Joyal
John I. Kirkpatrick
B. Alden Thresher
Robert J. Wert
Logan Wilson

ETS Officers

Henry Chauncey, *President*
William W. Turnbull, *Executive Vice President*
C. Russell de Burlo, Jr., *Vice President*
Henry S. Dyer, *Vice President*
John S. Helmick, *Vice President*
Robert J. Solomon, *Vice President*
G. Dykeman Sterling,
Vice President and Treasurer
Catherine G. Sharp, *Secretary*
Joseph E. Terral, *Assistant Vice President*
David J. Brodsky, *Assistant Treasurer*
John Graham, *Assistant Treasurer*

Foreword

The 1963 Invitational Conference on Testing Problems centered upon both theoretical and practical aspects of measurement. Speakers reviewed and analyzed the current thinking that underlies the basic concepts of norms, reliability, and validity. The potential of cognitive and non-cognitive tests was explored as were the social consequences of tests in general. In contrast to these theoretical discussions, the Conference featured two interesting reports on the application of objective tests in the field of medical education. All in all, it was a most stimulating program that balanced the reality of the present with implications for the future.

I should like to extend our thanks to Dr. Alexander G. Wesman who, as Chairman, was responsible for planning this program. We owe our thanks also to Dr. Jerome S. Bruner for his luncheon address and to the other distinguished speakers whose efforts made this Conference such a success.

Henry Chauncey
PRESIDENT

6

Preface

To be designated as chairman of the ETS Invitational Conference on Testing Problems is simultaneously an honor and an opportunity. The list of prior chairmen is a highly distinguished one; the excellence of previous programs is documented by the constancy of increase in attendance at the meetings. Opportunity is provided by the freedom given the chairman to compose the program as he wishes and by the regard in which the conference is held—a regard which predisposes desired speakers to accept the chairman's invitation. The chairman who fails to provide a stimulating meeting has only himself to blame; if he chooses wisely, the speakers will fulfill his responsibilities to his credit.

In organizing the 1963 conference, my design was to have basic concepts and concerns in the field of measurement presented comprehensively, informedly, and informatively. The first session was devoted to "state-of-the-science" overviews of three fundamental concepts—norms, reliability, and validity. Dr. Roger T. Lennon called for more active attention to development of norming theory, summarized current norming practices, and urged the establishment of a system which would permit comparable norms for tests whose primary standardizations are based on somewhat differing samples of the population. Dr. Robert L. Thorndike commented on proposals, practices, and procedures for estimating reliability; he structured his discussion in terms of concept formulations, construction of mathematical models, and methods of obtaining pertinent data. The third fundamental test characteristic, validity, was discussed by Dr. Anne Anastasi. Under the topic headings of construct validation, decision theory, moderator variables, synthetic validity, and response styles, she reviewed new approaches devised for the study of validity during the last ten years.

The second morning session was devoted to the report and appraisal of test use in a specific field of application—medicine. Dr. John P. Hubbard described a testing method used by the National Board of Medical Examiners to appraise the diagnostic competence of an intern, employing a sequential, programmed pattern in a realistic clinical situation. Under the title "Alternate Criteria in Medical Education and their Correlates," Dr. E. Lowell Kelly reported an extensive investigation of predictor and criterion variables of concern to medical schools.

At the luncheon meeting we were privileged to hear a most interesting address by Dr. Jerome S. Bruner. His discussion of learning processes and concept formation development was truly a highlight—stimulating, scholarly, and informative.

The three afternoon speakers directed our attention to implications and consequences of measurement. The first, Dr. Warren G. Findley, discussed current theory and application in cognitive fields—the appraisal of ability. He reviewed our changing approaches to investigating the structure and organization of mental ability, and our contemporary methods of appraising achievement in schools and colleges. Non-cognitive aspects of student performance were treated by Dr. Samuel Messick, who considered the potential contribution of personality assessment techniques to the prediction of success of college students. He undertook to raise (and answer) questions as to scientific standards for evaluating personality devices, and ethical problems in the use of such devices in practical decision making. The final speaker of the day was Dr. Robert L. Ebel, whose topic was "The Social Consequences of Educational Testing." Dr. Ebel examined the charges recently aimed at testing by sympathetic and by antagonistic critics; subjected the charges to judicial consideration; accepted the validity of some, rejected the validity of others; and brought the several issues into saner perspective in concluding remarks on the social consequences of *not* testing.

I would be lacking in gratitude if I failed to express explicitly my deep appreciation to the committee of previous Invitational Conference chairmen which selected me, and to Educational Testing Service which sponsored the meeting and supplied professional and practical assistance at every stage of the development of the program. It was for me a most rewarding experience.

Alexander G. Wesman
CHAIRMAN

Contents

iii Foreword by Henry Chauncey

v Preface by Alexander G. Wesman

Session I: Basic Concepts in Measurement —1963

13 Norms, Roger T. Lennon, Test Department, Harcourt, Brace and World, Inc.

23 Reliability, Robert L. Thorndike, Teachers College, Columbia University

33 Some Current Developments in the Measurement and Interpretation of Test Validity, Anne Anastasi, Department of Psychology, Graduate School, Fordham University

Session II: Testing and the Medical Profession

49 Programmed Testing in the Examinations of the National Board of Medical Examiners, John P. Hubbard, National Board of Medical Examiners

64 Alternate Criteria in Medical Education and Their Correlates, E. Lowell Kelly, Bureau of Psychological Services, The University of Michigan

Contents continued

Luncheon Address

86 Growing, Jerome S. Bruner,
Center for Cognitive Studies
Harvard University

Session III: Implications and Consequenc of Measurement

101 Ability and Performance, Warren G. Findley,
College of Education,
The University of Georgia

110 Personality Measurement and College Performance, Samuel Messick, Educational Testing Service

130 The Social Consequences of Educational Testin
Robert L. Ebel, School for Advanced Studies,
College of Education,
Michigan State University

144 List of Conference Participants

Session I

Theme:
Basic Concepts in Measurement—1963

Norms: 1963

Roger T. Lennon,
Test Department,
Harcourt, Brace
and World, Inc.

Several months ago, your Chairman extended to me his invitation, which I was most pleased to accept, to take part in today's proceedings. He said that he would like me to talk about norms. "Al," I said, "that is a rather broad topic. Can you give me any hints as to which aspects of it you would suggest I concern myself with?" By dint of patient questioning I was able to elicit from him his hope that I would undertake a review of developments over the past 15 or 20 years in norming theory, norming technology and related areas, a survey of current practices with respect to norming varieties of types of tests, a critical analysis thereof, a prospectus for needed improvement, and perhaps a prediction of future developments in the realm of test norming—all, however, not to consume more than 20 or at the most 25 minutes. Then, like all good chairmen, he said, "But, of course, use your own discretion," neatly combining this passing of the buck with the subtle flattery of crediting me with possession of some discretion.

I have found it convenient to organize my remarks under two topics, which I shall refer to as norming theory, on the one hand, and norming technology, on the other.

As to norming theory, I shall have relatively little to say—and this for the best of reasons, namely, that the past decade has seen little development in this area. Where the literature abounds with theoretical treatment of validity and reliability, it is almost devoid of systematic treatment of norming; the words "norms" and "norming," for example, have not even appeared in the index of the *Annual Review of Psychology* for the past three years.

Indeed, some of you may even wonder what I have in mind when I speak of "norming theory." Surely, you will say, everyone knows what norms are and why we need them; what more is there to it than that? Perhaps I can make my meaning clear by recalling that the administration of a test to an individual or a group can, in most instances, be thought of as akin to the conduct of a scientific experiment. Performance on a test, when interpreted according to suitable norms, serves as evidence supportive or not supportive of a hypothesis: this pupil has or has not made progress in reading during the past school year; the group using this textbook has made significantly greater progress than comparable students spending the same amount of time on this subject; etc. Now the inferences or conclusions that are drawn from this experiment-like testing are obviously conditioned by attributes of the norming group; but we have little in the way of a body of general principles relating test interpretation to norm group characteristics, little spelling out of the relations between norms, let us say, and test validity, little *theory*, in a word, of norming. I shall go no further in developing this concept; for purposes of this paper, suffice it to report, as I did a moment ago, that the past decade has been productive of very little advance in this area.

But if it appears that the past decade has been disappointing with respect to advances in norming theory, the picture with respect to norming technology and current practice is a more encouraging one. I discern at least four lines of development:

1. Applications of sampling theory to test standardization, particularly as reflected in the work of Frederic Lord, have pointed the way to more efficient data-gathering designs.

2. We have added substantially to our knowledge about community and school system variables related to performance on achievement and general mental ability tests. Dr. Jack Merwin, some three years ago, reviewing the literature on community

and school characteristics related to test performance, found some eighty-odd relevant studies; a decade ago, there was scarcely a score. The work of Dr. Flanagan and his associates in Project Talent has already eventuated in a wealth of information about characteristics related to performance on various types of tests at the secondary school level, some corroborative of earlier findings, others raising questions about certain assumptions hitherto widely acted upon in definition of norming populations.

3. There is a general willingness on the part of the major test-making agencies to commit the resources required for adequate test standardization, at least with respect to their most important test series.

4. The major test publishers, several years ago, began to give serious consideration to the use of a common anchor test in norming their respective tests, as a device for heightening comparability among the norms. This enterprise has moved forward less rapidly than it should have, a state of affairs for which, I regret to say, I am as much responsible as any one individual.

By way of documenting these points, and as introduction to additional points that I shall make, I ask you to bear with me while I read to you excerpts from the descriptions of the standardization programs for six of the most widely used batteries of tests.

TEST A

"Basic procedure for ruling out bias was to select a stratified sample of communities on which to base the norms. Communities were stratified on a composite of factors which have been found to be related to the measured intelligence of children in the community. Each community which volunteered to serve in the normative testing was evaluated with respect to the factors of: 1) per cent of adult illiteracy; 2) number of professional workers per thousand; 3) per cent of home ownerships; 4) median home rental value. On the basis of a composite of these factors each community was classed as very high, high, average, low, or very low. All the pupils present in each grade in the community were to be tested . . ."

TEST B

"Schools in the norm sample were so chosen that the representation from each of nine regions is similar to the proportions in the United States. At the inception of the program a random sample of all school superintendents in the country was chosen. The superintendents were asked if they were willing to participate in a long-range standardization program. The selection of schools was then random from all available schools in the region."

TEST C

"More important than the sheer number of students tested, however, is the degree to which they adequately represent the total national public school population at those grades. U. S. school enrollment data were obtained showing distributions of students by geographic region. Apportionment according to community size within each geographic region was based on 1960 census figures for the distribution of population among communities of various sizes. Invitations to participate in the standardization program were then extended to appropriate school systems, so selected that the group as a whole would typify the national population. Eighty-five school systems in thirty-seven states participated in the standardization program. All cooperating school systems were asked to test complete classroom groups from one or more schools so chosen as to be representative of the community."

TEST D

"The total pupil enrollment in public elementary and secondary schools in the United States is the reference population on which the norms are based . . . Data in the Biennial Survey of Education and general educational, social, cultural and economic conditions were considered in grouping states with similar characteristics into geographical regions. Specific characteristics considered were average expenditures per pupil for instructional purposes, length of school term, and type of school organization. Community size was the second factor used for stratification control. The norming samples for all grades within a given level were independent. Thus, any single school contributed to only one grade for any single level of the test. No one school was permitted to contribute to samples for two successive grades, even though they were for different levels of the test. A total of 672 school systems were contacted, of which 341 agreed and actually did participate in the norming program. A total of 69,345 pupils in 48 states were tested in this program."

TEST E

"The norms purport to describe the achievement of pupils 'representative' of the nation's public school population. Authors and publishers sought to obtain a norm group that would match the national school population with respect to certain characteristics known or assumed to be related to achievement. These characteristics include size of school system, geographical location, type of community, intelligence level of pupils and type of system (segregated or non-segregated). Each field representative was asked to designate 20 school systems meeting specifications that would yield a properly representative total norm group. A total of 225 systems accepted the invitation and carried through all necessary phases of the program. Included in this group are public school systems from 49 states; the number of pupils tested in the standardization program was over 500,000. One additional control relating to age was exercised in the selection of the final norm group. Pupils falling outside the 18 months range modal or typical for each grade were excluded from the norm group; the per cent of pupils thus excluded ranged from 10 to 20. Participating systems were required to test entire enrollments in at least three consecutive grades."

TEST F

"The population to which the norms apply includes all students in grades 9 through 12 in regular daily attendance at public high schools throughout the United States. The sample on which the norms are based was drawn so as to reflect the regional distribution and the community-size distributions for the national population. A preliminary sample of school systems was chosen strictly at random from each of the 36 strata. The number of systems chosen from each stratum was based on the average high school enrollment per grade within that stratum. This preliminary sample of 714 systems included approximately three times as many students as were demanded by the sample specifications. Invitations were issued to these 714 school systems, and over 200 school systems responded affirmatively. In multiple-building systems either all buildings or randomly selected buildings were included in the sample. All pupils in all grades in the cooperating schools were tested. A total of 366 schools in 254 school systems participated in the standardization project."

I do not cite these particular standardization projects as examples of either good or bad practice in norming tests; much less do I propose to criticize any features of any one of these programs. I adduce them rather as representative of current practice on the part of major test publishers with respect to standardization of their more important test offerings. The six excerpts are, by intent, chosen from publications of the six major test publishers; the excerpts are mostly verbatim, but not complete; three are for achievement batteries, three for general ability tests.

It seems to me quite clear from the descriptions that the norming in each of the instances cited must be judged to be a planful, earnest and informed attempt on the part of the respective authors and publishers to develop appropriate norms, an effort implying in every instance substantial commitment of time and resources. I may observe in passing that the author-publisher expenditure is likely to be in excess of 40 or 50 cents for each case tested in the standardization of a group (and very much higher in the case of an individual) test and you can readily appreciate the size of the commitment in these norming programs, involving as they did tens or even hundreds of thousands of pupils. It is no longer possible to say, as it might have been 20 years ago, that the norms represent adventitious collections of available test scores bearing only accidental relationship to an accurate description of the test performance of definable groups of pupils. At least with respect to the tests involved here—and they would collectively represent a large fraction of the testing done in elementary and secondary schools—such shortcomings as the norms may possess, either viewed individually or in relation to one another, do not stem from carelessness, lack of sophistication, or unwillingness to devote the resources needed to do respectable norming.

But shortcomings are in evidence; the norms do leave much to be desired, at least when viewed across tests. Threr are discernible marked differences with respect to the population whose achievement or ability the norms purport to describe; the variables considered important as stratifying variables; sampling procedures; the proportions of voluntary cooperation forthcoming; the degree of control over administration and scoring; and other critical characteristics. It is impossible to state on *a priori* grounds the effect that such differences may have in introducing systematic variations among the several sets of norms, but there are good reasons for supposing that the differences in norms ascribable simply to these variations in norming procedures are not negligible. When we consider that to such differences from test to test there must be added differences associated with varying content, with the time at which standardization programs are conducted, including the time of the school year, the issue of comparability, or lack of it, among the results of the various tests may begin to be seen in proper perspective. Empirical data reveal that there may be variations of as much as a year and a half in grade equivalent among the results yielded by various achievement tests; variations of as much as 8 or 10 points of IQ among various intelligence tests are, of course, by no means uncommon.

Some of you may feel that this lack of comparability among results of various tests is not really a matter of great concern —that as long as a school or school system consistently uses a given test or test series, it need not be too distressed that some other test or series would yield somewhat different results. If there be any such among you, may I cite for you a situation presently prevailing in the state of California, to the distress of both California educators and the test publishers. The California legislature, in response to public clamor over the quality of education in that state, enacted legislation prescribing the administration of ability and achievement tests in grades 5, 8, and 11, on an annual basis, to all public school pupils in the state. The state education department issued implementing regulations, which, in a wholly laudable attempt to provide for a measure of local autonomy in the selection of evaluation instruments, established an approved list of about half a dozen ability tests and an equal number of achievement tests from which local school districts might choose the instruments to be used. School districts are required to submit results to the state education department, which, in turn, is charged with the responsibility of preparing a summary of pupil achievement for the state for submission to the state board of education and presumably to the legislature and public. Now imagine the task that confronts the state education department in attempting to combine into a single summary the results, non-comparable as they are known to be, from a variety of tests. How can this agency discharge this responsibility and give to the legislature and the public a clear picture of pupil attainments? Must it undertake its own study of equivalence among the half dozen or so measures? This is an expensive and complicated undertaking, the results of which would in any case be subject to serious limitations. Must it resort to the alternative of requiring use of the same instrument by all school districts? I for one could consider this to be undesirable on various educational grounds.

Is this a state of affairs that we in testing should be willing to accept with complacency? I do not believe we should; I do not believe we have to. To do so, in my opinion, is to court growing lay and professional disbelief that measurement can provide worthwhile answers to important educational questions. Neither do I think, as do some within and without the testing field, that these vexing problems of norming should prompt us to repudiate the notion of national norms as an unattainable, unrealistic, and meaningless goal. For both general mental ability or scholastic aptitude measures and for achievement tests, there is surely a place and a need for a single, comprehensively based, broadly descriptive set of norms, whatever additional needs may also exist for data descriptive of particular samples of the general population. Rather, the proper direction for us now to take would seem to me to be along the path of a collaborative attack on the norming problem by the major test-producing agencies. I think each of us publishers should be willing to sacrifice whatever competitive advantage one or another of us may have felt he enjoyed by virtue of the superior norming of his tests, for the sake of the great gains in test interpretation that would flow from adoption of common definitions of norms populations and norming methods. We might even succeed in having schools give pre-eminence in selecting tests to considerations of content, validity and reliability.

Exactly 23 years ago this very day, speaking in this very forum, Dr. Cureton read a paper on norms that has not, in my opinion, been surpassed by any subsequent paper on this issue. Cureton called for a general adoption by test-making agencies of a system of anchoring their respective tests to a common scale. He urged the development of a basic anchor test, its standardization on a genuinely representative sample of the general population, and the equating of intelligence tests and achievement tests of all publishers to this common scale. The attainment of this state of affairs would mark, in Cureton's words, "the date of maturity of educational and mental measurements as a science, and of educational guidance and counseling as a profession." We are, alas, not yet at this level of maturity, by Cureton's definition.

While I have no reason to suppose that any appeal that I might make along these lines will be more potent than Dr.

Cureton's (save only that the need for some such development is now far more evident than it was in 1940), I would like to close my remarks with a similar call to concerted action now by the major test-making agencies. Surely we now know enough about the characteristics of communities and school systems related to performance on achievement and mental ability measures, and are sufficiently close to common understanding of the proper general population on which to develop norms, to enable us to agree on a generally acceptable definition of the population whose test performance we seek to describe; to specify the distribution of this population on measures of economic status, cultural status, educational effort and caliber of pupil population, plus other demographic features to which norming samples will be made to conform; to push ahead with the creation of an anchor instrument that will serve as a defining variable for all standardization groups at least for tests in the general cognitive domain, and thus to bring our collective efforts to that level of maturity for which Dr. Cureton pleaded. As we value the concept of a science of measurement of human abilities, let us take at least these steps to make our efforts more deserving of the label "scientific."

Reliability

ROBERT L. THORNDIKE,
*Teachers College,
Columbia University*

It is just 17 years ago that I had the honor of addressing this august assemblage – somewhat smaller and less imposing then than now – on "Logical Dilemmas in the Estimation of Reliability." I should have stopped when I was ahead! But some evil genie brought his power to bear upon your program chairman for this year, and here I am let out of the bottle again to comment on developments that have occurred in thinking about and dealing with the topic of reliability over the time span since last I held forth. I don't know whether I am being used as a practical example of the importance of test-retest reliability or as a demonstration of the fact that once ability to read statistical exposition has reached a maximum in the late twenties it goes into a positively accelerated curve of decline from that time on. Fortunately, few of you in this room today have any recollection of what I said in 1946 – or are likely to remember beyond your second cocktail this afternoon what I say today. Unfortunately, my deathless prose will be preserved for posterity in the *Proceedings* of the occasion. But for this there is no antidote.

The issues of test reliability may be approached, it has seemed to me, at three levels. The first of these is the verbal level of formulation and definition of the concept. A second level is that of mathematical model-building, leading to specification of a set of formulas and computational procedures by which the parameters specified in the model are to be estimated. A third level is that of experimental data-gathering procedures, under which certain tests are given to certain subjects at certain times and treated in certain ways to yield scores that are the raw materials to which we apply our formulas and computational procedures.

Developments in the past 17 years appear to have been primarily at the first two of these levels. In fact, Oscar Buros, addressing the American Educational Research Association last year, expressed the view that the last 35 years have been retrogressive, so far as our empirical procedures for appraising reliability are concerned. He exhorted us to return to the virtuous ways of our forefathers and stick to the operation of testing the individual with two or more experimentally independent tests, in order to get the data which permit generalizations about precision of measurement over occasions as well as over test items, and to this I can only say "Amen". He urged us not to backslide from the high standards of precision that Truman Kelley laid down for us in 1927, and to this I would comment "It all depends." But my point is that I am not aware of any distinctive proposals for new patterns of data-gathering that call for our special attention today, though it is always well that we be aware of the limitations of the methods we are using.

Turning now to verbal formulation, perhaps the major trend has been toward increasingly explicit formulation of the concept that performance on a test should be thought of as a sample from a defined universe of events, and that reliability is concerned with the precision with which the test score, that is, the sample, represents the universe. I shall not try to be a historian, but will merely note that this idea has been made fairly explicit by Buros, by Cronbach, by Tryon, and probably by others.

What we may call the "classical" approach to reliability tended to be conceptualized in terms of some unobservable underlying "true score" distorted in a given measurement by an equally unobservable "error of measurement." The corresponding mathematical models and computational routines were procedures for estimating the magnitude, absolute or relative, of this measurement error. The formulation in terms of sampling does away in one lightning stroke with the mystical "true score," somehow enshrined far above the mundane world of scores and data, and replaces it with the less austere "expected value" of the score in the population of values from which the sample score was drawn.

Now what are the implications, the advantages, and possibly the limitations of this "sampling" conception over the classic "true score and error" conception?

For myself, I cannot say that the advantage lies in simplification and clarification. This notion of a "universe of possible scores" is in many ways a puzzling and somewhat confusing one. Of what is this universe composed? Suppose we have given Form A of the XYZ Reading Test to the fifth graders in our school and gotten a score for each pupil. Of what universe of scores are these scores a sample—of all possible scores that we might have gotten by giving Form A on that day? Of all possible scores that we might have gotten by giving Form A sometime that month? Of all possible scores that might have been gotten by giving Forms A or B or C or other forms up to a still-unwritten Form K on that day? Of scores on these same numerous and presumably "parallel" forms—and we shall have to ask what "parallel" means under a sampling conception of reliability—at some unspecified date within the month? Of scores on the whole array of different reading tests produced by different authors over the past 25 years? Of scores on tests of some aspect of educational achievement not further specified?

As soon as we try to conceptualize a test score as a sample from some universe, we are brought face to face with the very knotty problem of defining the universe from which we are sampling. But I suppose this very difficulty may be in one sense a blessing. The experimental data-gathering phase of estimating reliability has *always* implied a universe to which those data corresponded. Split-half procedures refer only to a universe of behaviors produced at one single point in time, retest procedures to a universe of responses to a specific set of items, and so forth. Perhaps one of the advantages of the sampling formulation is that it makes us more explicitly aware of the need to define the universe in which we are interested, or to acknowledge the universe to which our data apply. Certainly, over the past

30 years, all of us who have written for students and for the test-using public have insistently harped upon the nonequivalence of different operations for estimating reliability, and emphasized the different universes to which different procedures referred.

The notion of a random sample from a universe of responses seems most satisfying and clear-cut when we are dealing with some unitary act of behavior, which we score in some way. Examples would be distance jumped in a broad jump, time to run 100 yards, speed of response on a trial with a reaction time device, or number of trials to learn a series of nonsense syllables to a specified level of mastery. In these cases, the experimental specification of the task is fairly complete. Thus, for the 100 yard run, we specify a smooth, straight, well-packed cinder track, a certain type of starting blocks, certain limitations on the shoes to be worn, a certain pattern of preparatory and starting signals, and a certain procedure for recording time. A universe could then be the universe of times for a given runner, over a certain span of days, weeks or months of his running career. Data from two or more trials under these conditions would give us some basis for generalizing about the consistency of this behavior for this defined universe. We could also extend the universe if we wished—to include wooden indoor tracks for example, or to include running on grass, or running in sneakers instead of track shoes—and sample randomly from this more varied universe. As conditions were varied, we might expect typical performance to vary more widely and precision to be decreased.

We are usually interested in estimating precision for each of a population of persons, rather than just for some one specific person, and so we are likely to have a sampling from some population of persons. The nature of that population will also influence estimates of precision, and so it will be important that the population be specified as well as the conditions. Precision of estimating time to run 100 yards is probably much greater for college track stars than for middle-aged professors—for whom one might occasionally get scores approaching infinity.

But it would be possible to specify the population of individuals fairly satisfactorily, as well as the population of behaviors for a given person. Within this at least two-dimensional universe, we could sample in a presumably random fashion, and we could then analyze our sample of observations to yield estimates of the relative precision with which a person could be located within the group or the absolute precision with which his time could be estimated in seconds.

When we are dealing with the typical aptitude or achievement test, however, in which the score is some type of summation of scores upon single items, the conception of the universe from which we have drawn a sample becomes a little more fuzzy. Here, fairly clearly, we are concerned with a sampling not only of responses to a given situation but also of situations to be responded to. How shall we define that universe? The classical approach to reliability tended to deal with this issue by postulating a universe of equivalent or parallel tests and by limiting the universe from which our sample is drawn to this universe of parallel tests. Parallel tests may be defined statistically as those having equal means, standard deviations, and correlations with each other and with other variables. But they may also be defined in terms of the operations of construction, as tests built by the same procedures to the same specifications. If we adopt the second definition, statistical characteristics will not be identical, but the tests will vary in their statistical attributes to the extent that different samples of items all chosen to conform to a uniform blueprint or test plan will produce tests with somewhat differing statistical values.

But some of the recent discussions seem to imply a random sampling of tests from some rather loosely and broadly defined domain—the domain of scholastic aptitude tests, or the domain of reading comprehension tests, or the domain of personal adjustment inventories. Clearly, these are very vague and ill-defined domains. A sampling expert would be hard put to delimit the universe or to propose any meaningful set of operations for sampling from it. And in the realm of practical politics, I question whether anyone has ever seriously undertaken to carry out such a sampling operation. One might argue that the data appearing in the manual of the XYZ Reading Test, showing its correlations with other published reading tests, are an approximation of such a domain sampling. But how truly do the set of tests, taken collectively, represent a random sampling from the whole domain of reading tests? One suspects that the tests selected for correlating were chosen by the author or publisher on some systematic and non-random basis—because they were widely used tests, because data with respect to them were readily available, or for some other non-random reason.

We note, further, that as we broaden our conception of the universe being sampled from that of all tests made to a certain uniform set of specifications to all tests of a certain ability or personality domain, we begin to face the issue of whether we are still getting evidence on reliability or whether we are now getting evidence on some aspect of construct validity. But, once again, perhaps we should consider it a contribution of the sampling approach that it makes explicit to us and heightens our awareness of the continuity from reliability to validity. Cronbach offers the single term "generalizability" to cover the whole gamut of relationships from those within the most restricted universe of near-exact replications to those extending over the most general and broadly defined domain and develops a common statistical framework which he applies to the whole gamut. Recognition that the same pattern of statistical analysis can be used whether one is dealing with the little central core, or with all the layers of the whole onion may be useful. On the other hand, we may perhaps question whether this approach helps to clarify our meaning of "reliability" as a distinctive concept.

A third context in which the random sampling notion has been applied to the conceptualization of reliability has been the context of the single test item. That is, one can conceive of a certain universe of test items—let us say the universe of vocabulary items, for example. A given test may be considered to represent a random sampling drawn from this item universe. This conception provides the foundation for the estimation of test reliability from the interrelations of the items of the sample, and thus to a somewhat more generalized and less restrictive form of the Kuder-Richardson reliability estimates.

But here, again, we encounter certain difficulties. These center on the one hand upon the definition of the universe and on the other upon the notion of randomness in sampling. In the first place, there are very definite constraints upon the items which make up our operational, as opposed to a purely hypothetical, universe. If we take the domain of vocabulary items as our example, we can specify what some of these constraints might be in an actual case. Firstly, there is typically a constraint upon the format of the item—most often to a 5-choice multiple choice form. Secondly, there are constraints imposed by editorial policy—exemplified by the decision to exclude proper names or specialized technical terms, or by a requirement that the options call for gross rather than fine discriminations of shade of meaning. Thirdly, there are the constraints that arise out of the particular idiosyncrasies of the item writers: their tendency to favor particular types of words, or particular tricks of mislead construction. Finally, there are the constraints imposed by the item selection procedures—selection to provide a predetermined spread of item difficulties and to eliminate items failing to discriminate at a designated level. Thus, the universe is considerably restricted, is hard to define, and the sampling from it is hardly to be considered random.

Presumably we could elaborate and delimit more fully the definition of the universe of items. Certainly, we could replace the concept of random sampling with one of stratified sampling, and indeed Cronbach has proposed that the sampling concept be extended to one of stratified sampling. But we may find that a really adequate definition of the universe from which we have sampled will become so involved as to be meaningless. We will almost certainly find that in proportion as we provide detailed specifications for stratification of our universe of items, and carry out our sampling within such strata, we are once again getting very close to a bill of particulars for equivalent tests. Just as random sampling is less efficient than stratified sampling, in opinion surveys or demographic studies, when stratification is upon relevant variables, so also random sampling of test items is less efficient than stratified sampling in making equivalent tests. Analytical techniques developed on the basis of random sampling assumptions will make a test appear less precise than it is as a representation of a population of tests which sample in a uniform way from different strata of the universe of items. It is partly in this sense that the Kuder-Richardson Formula 20 and the other formulas that try to estimate test reliability from item data, or from such test statistics as means and variances (which grow out of item data), are lower bound estimates of reliability. They treat the sampling of items as random rather than stratified. They assume that differences in item factor composition either do not exist, or are only such as arise by chance.

Sometimes the facts suggest that this may be approximately the case. Thus, Cronbach compared the values that he obtained for tests divided into random halves and those divided into judgmentally equivalent halves for a mechanical reasoning test, and found an average value of .810 for random splits and .820 for parallel splits. For a short morale scale the corresponding values were .715 and .737. But frequently a test is fairly sharply stratified—by difficulty level, by area of content, by intellectual process. When this is true, correlation estimates based on random sampling concepts may seriously underestimate those that would be obtained between two parallel forms of the test, and consequently the precision with which a given test represents the stratified universe.

These reactions to random sampling as applied to tests and test items were stimulated in part by Dr. Loevinger's presidential address to Division 5 at the recent APA meetings, and I gladly acknowledge the indebtedness, without holding her responsible for anything silly that I have said.

The shift in verbal formulation to a sampling formulation is compatible with a shift in mathematical models of reliability to analysis of variance and intraclass correlation models. These models have, of course, been proposed for more than 20 years, but they have been more systematically and completely expressed in the past decade.

The most comprehensive and systematic elaboration of this formulation of which I am aware is the one which has been distributed in rexographed form by Oscar Buros, and which is available for $1.00 from the Gryphon Press. I confess my own limitation when I say that I find this presentation pretty hard to follow. Hopefully, others of you will be either more familiar with or more facile at picking up the notation that Buros has used, and will be able to pick from the host of formulas that are offered the one that is appropriate to the specific data with which you are faced.

One great virtue of analysis of variance models is their built-in versatility. They can handle item responses that are scored 0 or 1, trial scores that yield scores with some type of continuous distribution, or, where more than one test has been given to each individual, scores for total tests. They can deal with the situation in which the data for each individual are generated by the same test or the same rater and also the situation in which test or rater vary from person to person. This latter situation is one of very real importance in many practical circumstances. How shall we judge the precision of a reported IQ when we do not know which of two or more forms of a test was given? How shall we appraise the repeatability of a course grade when the grade may be given by any one of the several different instructors who handle a course? If we have more than a single score for each individual, even though the scores are based on different tests or raters for each individual, we can get an estimate of within-persons variation. And whenever we have an estimate of within-persons variation we have a basis for judging the precision of a score or rating as describing a person. Clearly, with only two or three or four scores per person, the estimate of within-persons variance is very crude for a single individual. We must be willing to assume that the within-persons variance is sufficiently uniform from person to person for a pooling of data over persons to give us a usable common estimate of variance from test to test for each single individual. Having such an estimate, we can express reliability either as the precision of a score for an individual stated in absolute terms or as the precision of placement of an individual relative to his fellows.

As various writers have shown, the conventional Kuder-Richardson formulas emerge as special cases of the more general variance analysis approach. Likewise, the adjustment of correlational measures of reliability for test length are derivable from general variance analysis formulas.

I shall not try to recite to you a set of formulas today, because this would serve no good purpose. Rather, let me direct you to Tryon's 1957 article in the *Psychological Bulletin*, Buros' available if unpublished material, and Cronbach's forthcoming article in the *British Journal of Statistical Psychology*. These, plus Horsts' and Ebel's articles in *Psychometrika* should give you all the formulas you can use.

In closing, let me raise with you the question of how much you are willing to pay for precision in a given measurement. The cost is partly one of time and expense. But, given some fixed limit on time and expense, the cost can then be a cost in scope and comprehensiveness. We can usually make gains in precision by increasing the redundance and repetitiveness of successive observations. The more narrowly a universe is defined, the more adequately a given length of test sample can represent it. With all due respect to the error of measurement, we must recognize that it is often the error of estimate that we are really interested in. To maximize prediction of socially useful events, it may be advantageous to sacrifice a little precision in order to gain a greater amount of scope. Precision and high reliability are, after all, a means rather than an end.

Some Current Developments in the Measurement and Interpretation of Test Validity

Anne Anastasi,
Department of Psychology,
Graduate School,
Fordham University

Within the past decade, psychologists have been especially active in devising novel and imaginative approaches to test validity. In the time allotted, I can do no more than whet your appetite for these exciting developments and hope that you will be stimulated to examine the sources cited for an adequate exposition of each topic. I have selected five developments to bring to your attention. Ranging in scope from broad frameworks to specific techniques and from highly theoretical to immediately practical, these topics pertain to: construct validation, decision theory, moderator variables, synthetic validity, and response styles.

Construct Validation

It is nearly ten years since the American Psychological Association published its *Technical Recommendations* (1) outlining four types of validity: content, predictive, concurrent, and construct. As the most complex, inclusive, and controversial of the four, construct validity has received the greatest attention during the subsequent decade. When first proposed in the *Technical Recommendations*, construct validation was characterized as a validation of the theory underlying a test. On the basis of such a theory, specific hypotheses are formulated regarding the expected variations in test scores among individuals or among conditions, and data are then gathered to test these hypotheses. The constructs in construct validity refer to postulated attributes or traits that are presumably reflected in test performance. Concerned with a more comprehensive and more abstract kind of behavioral description than those provided by other types of validation, construct validation calls for a continuing accumulation of information from a variety of sources. Any data throwing light on the nature of the trait under consideration and the conditions affecting its development and manifestations contribute to the process of construct validation. Examples of relevant procedures include checking an intelligence test for the anticipated increase in score with age during childhood, investigating the effects of experimental variables such as stress upon test scores, and factor analyzing the test along with other variables.

Subsequently the concept of construct validity has been attacked, clarified, elaborated, and illustrated in a number of thoughtful and provocative articles by Cronbach and Meehl (14), Loevinger (30), Bechtoldt (6), Jessor and Hammond (28), Campbell and Fiske (11), and Campbell (10). In the most recent of these papers, Campbell (10) integrates much that had previously been written about construct validity and gives a well-balanced presentation of its contributions, hazards, and common misunderstandings. Referring to the earlier paper prepared jointly with Fiske (11), Campbell again points out that in order to demonstrate construct validity we need to show not only that a test correlates highly with other variables with which it should correlate, but also that it does not correlate with variables from which it should differ. The former is described as convergent validation, the latter as discriminant validation.

In their multitrait-multimethod matrix, Campbell and Fiske (11) proposed a systematic experimental design for this twin approach to validation. Essentially what is required is the assessment of two or more traits by two or more methods. Under these conditions, the correlations of the same trait assessed by different methods represent a measure of convergent validity (these correlations should be high). The correlations of different traits assessed by the same or similar methods provide a measure of discriminant validity (these correlations should be low or negligible). In addition, the correlations of the same trait independently assessed by the same method give an index of reliability.

Without attempting an evaluation of construct validity, for which I would urge you to consult the sources cited, I should nevertheless like to make a few comments about it. First, the basic idea of construct validity is not new. Some of the earliest tests were designed to measure such theoretical constructs as attention and memory, not to mention that most notorious of constructs, "intelligence." On the other hand, construct validity has served to focus attention on the desirability of basing test construction on an explicitly recognized theoretical foundation. Both in devising a new test and in setting up procedures for its validation, the investigator is urged to formulate psychological hypotheses. The proponents of construct validity have thus tried to integrate psychological testing more closely with psychological theory and experimental methods.

With regard to specific validation procedures, construct validation also utilizes much that is not new. Age differentiation, factorial validity, and the effect of such experimental variables as practice on test scores have been reported in test manuals long before construct validity was given a name in the *Technical Recommendations.* As a matter of fact, the methodology of construct validity is so comprehensive as to encompass even the procedures characteristically associated with other types of validity (see 2, Ch. 6). Thus the correlation of a mechanical aptitude test with subsequent performance on engineering jobs would contribute to our understanding of the construct measured by this test. Similarly, comparing the performance of neurotics and normals is one way of checking the construct validity of a test designed to measure anxiety. Nevertheless, construct validation has stimulated the search for novel ways of gathering validation data. Although the principal techniques currently employed to investigate construct validity have long been familiar, the field of operation has been expanded to admit a wider variety of procedures.

The very multiplicity of data-gathering techniques recognized by construct validity presents certain hazards. As Campbell puts it, the wide diversity of acceptable validational evidence "makes possible a highly opportunistic selection of evidence and the editorial device of failing to mention validity probes that were not confirmatory" (10, p. 551). Another hazard stems from misunderstandings of such a broad and loosely defined concept as construct validity. Some test constructors apparently interpret construct validation to mean content validity expressed in terms of psychological trait names. Hence they present as construct validity purely subjective accounts of what they believe (or hope) their test measures.

It is also unfortunate that the chief exponents of construct validity stated in one of their articles that this type of validation "is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not 'operationally defined'" (14, p. 282). Such an assertion opens the door wider for subjective claims and fuzzy thinking about test scores and the traits they measure. Actually the theoretical construct or trait assessed by any test can be defined in terms of the operations performed in establishing the validity of the test. Such a definition should take into account the various external criteria with which the test correlated significantly, as well as the conditions that affect its scores. These procedures are entirely in accord with the positive contributions of construct validity. It would also seem desirable to retain the concept of criterion in construct validation, not as a specific practical achievement to be predicted, but as a general name for independently gathered external data. The need to base all validation on data rather than on armchair speculation would thus be re-emphasized, as would the need for data external to the test scores themselves.

Decision Theory

Even broader than construct validity in its scope and implications is the application of decision theory to test construction and evaluation (see 2, Ch. 7; 13; 25). Because of many technical complexities, however, the current impact of decision theory on test development and use is limited and progress has been slow.

Statistical decision theory was developed by Wald (37) with special reference to the decisions required in the inspection and quality control of industrial products. Many of its possible implications for psychological testing have been systematically worked out by Cronbach and Gleser in their 1957 book on *Psychological Tests and Personnel Decisions* (13). Essentially, decision theory is an attempt to put the decision-making process into mathematical form, so that available information may be used to reach the most effective decisions under specified circumstances. The mathematical procedures required by decision theory are often quite complex, and few are in a form permitting their immediate application to practical testing problems. Some of the basic concepts of decision theory, however, can help in the reformulation and clarification of certain questions about tests.

A few of these concepts were introduced in psychological testing before the formal development of statistical decision theory and were later recognized as fitting into that framework. One example is provided by the well-known Taylor-Russell Tables (36), which permit an estimate of the net gain in selection accuracy attributable to the use of a test. The information required for this purpose includes the validity coefficient of the test, the selection ratio, and the proportion of successful applicants selected without the use of the test. The rise in proportion of successful applicants to be expected from the introduction of the test is taken as an index of the test's effectiveness.

In many situations, what is wanted is an estimate of the effect of the test, not on proportion of persons who exceed the minimum performance, but on the over-all output of the selected group. How does the level of criterion achievement of the persons selected on the basis of the test compare with that of the total applicant sample that would have been selected without the test? Following the work of Taylor and Russell, several investigators addressed themselves to this question. It was Brogden (8) who first demonstrated that the expected increase in output or achievement is directly proportional to the validity of the test. Doubling the validity of the test will double the improvement in output expected from its use. Following a similar approach (see 27), Brown and Ghiselli (9) prepared a table whereby mean standard criterion score of the selected group can be estimated from a knowledge of test validity and selection ratio.

Decision theory incorporates a number of parameters not traditionally considered in evaluating the predictive effectiveness of tests. The previously mentioned selection ratio is one such parameter. Another is the cost of administering the testing program. Thus a test of low validity would be more likely to be retained if it were short, inexpensive, adapted for group administration, and easy to give. An individual test requiring a trained examiner and expensive equipment would need a higher validity to justify its retention. A further consideration is whether the test measures an area of criterion-relevant behavior not covered by other available techniques.

Another major aspect of decision theory pertains to the evaluation of outcomes. The absence of adequate systems for assigning values to outcomes is one of the principal obstacles in the way of applying decision theory. It should be noted, however, that decision theory did not introduce the problem of values into the decision process, but merely made it explicit. Value systems have always entered into decisions, but they were not heretofore clearly recognized or systematically handled.

Still another feature of decision theory is that it permits a consideration of the interaction of different variables. An example would be the interaction of applicant aptitudes with alternative treatments, such as types of training programs to which individuals could be assigned. Such differential treatment would further improve the outcome of decisions based on test scores. Decision theory also focuses attention on the important fact that the effectiveness of a test for selection, placement, classification, or any other purpose must be compared not with chance or with perfect prediction, but with the effectiveness of other available predictors. The question of the base rate is also relevant here (33). The examples cited provide a few glimpses into ways in which the application of decision theory may eventually affect the interpretation of test validity.

Moderator Variables

A promising recent development in the interpretation of test validity centers on the use of moderator variables (7, 19, 21, 22, 23, 24, 35). The validity of a given test may vary among subgroups or individuals within a population. Essentially the problem of moderator variables is that of predicting these differences in predictability. In any bivariate distribution, some individuals fall close to the regression line, others miss it by appreciable distances. We may then ask whether there is any characteristic in which those falling farther from the regression line, for whom prediction errors are large, differ systematically and consistently from those falling close to it. Thus a test might be a better predictor of criterion performance for men than for women, or for applicants from a lower than for applicants from a higher socioeconomic level. In such examples, sex and socioeconomic level are the moderator variables since they modify the predictive validity of the test.

Even when a test is equally valid for all subgroups, the same score may have a different predictive meaning when obtained by members of different subgroups. For example, if two students with different educational backgrounds obtain the same score on the Scholastic Aptitude Test, will they do equally well in college? Or will the one with the poorer or the one with the better background excel? Moderator variables may thus influence cutoff scores, regression equation weights, or validity coefficients of the same test for different subgroups of a population.

Interests and motivation often function as moderator variables in individual cases. If an applicant has little interest in a job, he will probably do poorly regardless of his scores on relevant aptitude tests. Among such persons, the correlation between aptitude test scores and job performance would be low. For individuals who are interested and highly motivated, on the other hand, the correlation between aptitude test score and job success may be quite high. From another angle, personality inventories like the MMPI may have higher validity for some types of neurotics than for others (19). The characteristic behavior of the two types may make one more careful and accurate in reporting symptoms, the other careless or evasive.

A moderator variable may itself be a test score, in terms of which individuals may be sorted into subgroups. There have been some promising attempts to identify such moderator variables in test scores (7, 19, 21, 24). In a study of taxi drivers conducted by Ghiselli (21), the correlation between an aptitude test and a criterion of job performance was only .22. The group was then sorted into thirds on the basis of scores on an occupational interest inventory. When the validity of the aptitude test was recomputed within the third whose occupational interest level was most appropriate for the job, it rose to .66. Such findings suggest that one test might first be used to screen out individuals for whom the second test is likely to have low validity; then from among the remaining cases, those scoring high on the second test are selected.

Even within a single test, such as a personality inventory, it may prove possible to develop a moderator key in terms of which the validity of the rest of the test for each individual can be assessed (24). There is also evidence suggesting that intra-individual variability from one part of the test to another affects the predictive validity of a test for individuals (7). Those individuals for whom the test is more reliable (as indicated by low intra-individual variability) are also the individuals for whom it is more valid, as might be anticipated.

Synthetic Validity

A technique devised to meet a specific practical need is synthetic validity (4, 29, 34). It is well known that the same test may have high validity for predicting the performance of office clerks or machinists in one company and low or negligible validity for jobs bearing the same title in another company. Similar variation has been found in the correlations of tests with achievement in courses of the same name given in different colleges. The familiar criterion of "college success" is a notorious example of both complexity and heterogeneity. Although traditionally indentified with grade point average, college success can actually mean many different things, from being elected president of the student council or captain of the football team to receiving Phi Beta Kappa in one's junior year. Individual colleges vary in the relative weights they give to these different criteria of success.

It is abundantly clear that: (1) educational and vocational criteria are complex; (2) the various criterion elements, or subcriteria, for any given job, educational institution, course, etc. may have little relation to each other; and (3) different criterion situations bearing the same name often represent a different combination of sub-criteria. It is largely for these reasons that test users are generally urged to conduct their own local validation studies. In many situations, however, this practice may not be feasible for lack of time, facilities, or adequate samples. Under these circumstances, synthetic validity may provide a satisfactory approximation of test validity against a particular criterion. First proposed by Lawshe (29) for use in industry, synthetic validity has been employed chiefly with job criteria, but it is equally applicable to educational criteria.

In synthetic validity, each predictor is validated, not against a composite criterion, but against job elements identified through job analysis. The validity of any test for a given job is then computed synthetically from the weights of these elements in the job and in the test. Thus if a test has high validity in predicting performance in delicate manipulative tasks, and if such tasks loom large in a particular job, then the test will have high synthetic validity for that job. A statistical technique known as the J-coefficient (for Job-coefficient) has been developed by Primoff (34) for estimating the synthetic validity of a test. This technique offers a possible tool for generalizing validity data from one job or other criterion situation to another without actually conducting a separate validation study in each situation. The J-coefficient may also prove useful in ordinary battery construction as an intervening step between the job analysis and the assembling of a trial battery of tests. The preliminary selection of appropriate tests is now done largely on a subjective and unsystematic basis and might be improved through the utilization of such a technique as the J-coefficient.

Response Styles

The fifth and last development I should like to bring to your attention pertains to response styles. Although research on response styles has centered chiefly on personality inventories, the concept can be applied to any type of test. Interest in response styles was first stimulated by the identification of certain test-taking attitudes which might obscure or distort the traits that the test was designed to measure. Among the best known is the social desirability variable, extensively investigated by Edwards (15, 16, 17, 18). This is simply the tendency to choose socially desirable responses on personality inventories. To what extent this variable should also reflect the tendency to choose common responses is a matter on which different investigators disagree. Other examples of response styles include acquiescence, or the tendency to answer "yes" rather than "no" regardless of item content (3, 5, 12, 20); evasiveness, or the tendency to choose question marks or other indifferent responses; and the tendency to utilize extreme response categories.

We can recognize two stages in research on response styles. First there was the recognition that stylistic components of item form exert a significant influence upon test responses. In fact, a growing accumulation of evidence indicated that the principal factors measured by many self-report inventories were stylistic rather than content factors. At this stage, such stylistic variance was regarded as error variance, which would reduce test validity. Efforts were therefore made to rule out these stylistic factors through the reformulation of items, the development of special keys, or the application of correction formulas.

More recently there has been an increasing realization that response styles may be worth measuring in their own right. This point of view is clearly reflected in the reviews by Jackson and Messick (26) and by Wiggins (38), published within the past five years. Rather than being regarded as measurement errors to be eliminated, response styles are now being investigated as diagnostic indices of broad personality traits. The response style that an individual exhibits in taking a test may be associated with characteristic behavior he displays in other, nontest situations. Thus the tendency to mark socially desirable answers may be related to conformity and stereotyped conventionality. It has also been proposed that a moderate degree of this variable is associated with a mature, individualized self concept, while higher degrees are associated with intellectual and social immaturity (31, 32). With reference to acquiescence, there is some suggestive evidence that the predominant "yeasayers" tend to have weak ego controls and to accept impulses without reservation, while the predominate "naysayers" tend to inhibit and suppress impulses and to reject emotional stimuli (12).

The measurement of response styles may provide a means of capitalizing on what initially appeared to be the chief weaknesses of self-report inventories. Several puzzling and disappointing results obtained with personality inventories seem to make sense when reexamined in the light of recent research with response styles. Much more research is needed, however, before the measurement of response styles can be put to practical use. We need more information on the relationships among different response styles, such as social desirability and acquiescence, which are often confounded in existing scales. We also need to know more about the inter-relationships among different scales designed to measure the same response style. And above all, we need to know how these stylistic scales are related to external criterion data.

The five developments cited in this paper represent ongoing activities. It is premature to evaluate the contribution any of them will ultimately make to the measurement or interpretation of test validity. At this stage, they all bear watching and they warrant further exploration.

REFERENCES

1. American Psychological Association. *Technical Recommendations for Psychological Tests and Diagnostic Techniques.* Washington: American Psychological Association, 1954.
2. Anastasi, Anne. *Psychological Testing.* (2nd ed.) New York: Macmillan, 1961.

3. Asch, M. J. "Negative Response Bias and Personality Adjustment." *Journal of Counseling Psychology*, 1958, 5, 206-210.

4. Balma, N. J., Ghiselli, E. E., McCormick, E. J., Primoff, E. S., & Griffin, C. H. "The Development of Processes for Indirect or Synthetic Validity—A Symposium." *Personnel Psychology*, 1959, 12, 395-420.

5. Bass, B. M. "Authoritarianism or Acquiescence." *Journal of Abnormal and Social Psychology*, 1955, 51, 611-623.

6. Bechtoldt, H. P. "Construct Validity: A Critique." *American Psychologist*, 1959, 14, 619-629.

7. Berdie, R. F. "Intra-individual Variability and Predictability." *Educational and Psychological Measurement*, 1961, 21, 663-676.

8. Brogden, H. E, "On the Interpretation of the Correlation Coefficient as a Measure of Predictive Efficiency." *Journal of Educational Psychology*, 1946, 37, 65-76.

9. Brown, C. W., & Ghiselli, E. E. "Per Cent Increase in Proficiency Resulting From Use of Selective Devices." *Journal of Applied Psychology*, 1953, 37, 341-345.

10. Campbell, D. T. "Recommendations for APA Test Standards Regarding Construct, Trait, and Discriminant Validity." *American Psychologist*, 1960, 15, 546-553.

11. Campbell, D. T., & Fiske, D. W. "Convergent and Discriminant Validation by the Multitrait-multimethod Matrix." *Psychological Bulletin*, 1959, 56, 81-105.

12. Couch, A., & Keniston, K. "Yeasayers and Naysayers: Agreeing Response Set as a Personality Variable." *Journal of Abnormal and Social Psychology*, 1960, 60, 151-174.

13. Chronbach, L. J., & Gleser, Goldine C. *Psychological Tests and Personnel Decisions*. Urbana, Illinois: University of Illinois Press, 1957.

14. Cronbach, L. J., & Meehl, P. E. "Construct Validity in Psychological Tests." *Psychological Bulletin*, 1955, 52, 281-302.

15. Edwards, A. L. *The Social Desirability Variable in Personality Assessment and Research*. New York: Dryden, 1957.

16. Edwards, A. L., & Diers, Carol J. "Social Desirability and the Factorial Interpretation of the MMPI." *Educational and Psychological Measurement*, 1962, 22, 501-508.

17. Edwards, A. L., Diers, Carol J., & Walker, J. N. "Response sets and Factor Loadings on Sixty-one Personality Scales." *Journal of Applied Psychology*, 1962, 46, 220-225.

18. Edwards, A. L., & Heathers, Louise B. "The First Factor of the MMPI: Social Desirability or Ego Strength?" *Journal of Consulting Psychology*, 1962, 26, 99-100.

19. Fulkerson, S. C. "Individual Differences in Response Validity." *Journal of Clinical Psychology*, 1959, 15, 169-173.

20. Gage, N. L., Leavitt, G. S., & Stone, G. C. "The Psychological Meaning of Acquiescence Set for Authoritarianism". *Journal of Abnormal and Social Psychology*, 1957, 55, 98-103.

21. Ghiselli, E. E. "Differentiation of Individuals in Terms of Their Predictability." *Journal of Applied Psychology*, 1956, 40, 374-377.

22. Ghiselli, E. E. "Differentiation of Tests in Terms of the Accuracy With Which They Predict for a Given Individual." *Educational and Psychological Measurement*, 1960, 20, 675-684.

23. Ghiselli, E. E. "The Prediction of Predictability." *Education and Psychological Measurement*, 1960, 20, 3-8.

24. Ghiselli, E. E. "Moderating Effects and Differential Reliability and Validity." *Journal of Applied Psychology*, 1963, 47, 81-86.

25. Girshick, M. A. "An Elementary Survey of Statistical Decision Theory." *Review of Educational Research*, 1954, 24, 448-466.

26. Jackson, D. N., & Messick, S. "Content and Style in Personality Assessment." *Psychological Bulletin*, 1958, 55, 243-252.

27. Jarett, R. F. "Per Cent Increase in Output of Selected Personnel as an Index of Test Efficiency." *Journal of Applied Psychology*, 1948, 32, 135-145.

28. Jessor, R., & Hammond, K. R. "Construct Validity and the Taylor Anxiety Scale." *Psychological Bulletin*, 1957, 54, 161-170.

29. Lawshe, C. H., & Steinberg, M. D. "Studies in Synthetic Validity. 1. An Exploratory Investigation of Clerical Jobs." *Personnel Psychology*, 1955, 8, 291-301.

30. Loevinger, Jane. "Objective Tests as Instruments of Psychological Theory." *Psychological Reports*, 1957, 3, 635-694.

31. Loevinger, Jane. "A Theory of Test Response." *Proceedings of the* 1958 *Invitational Conference on Testing Problems*, Princeton, New Jersey: Educational Testing Service, 1959, 36-47.

32. Loevinger, Jane, & Ossorio, A. G. "Evaluation of Therapy by Self-Report: A Paradox." *American Psychologist*, 1958, 13, 366.

33. Meehl, P. E., & Rosen, A. "Antecedent Probability and the Efficiency of Psychometric Signs, Patterns, or Cutting Scores." *Psychological Bulletin*, 1955, 52, 194-216.

34. Primoff, E. S. "The J-coefficient Approach to Jobs and Tests." *Personnel Administration*, 1957, 20, 34-40.

35. Saunders, D. R. "Moderator Variables in Prediction." *Educational and Psychological Measurement*, 1956, 16, 209-222.

36. Taylor, H. C., & Russell, J. T. "The Relationship of Validity Coefficients to the Practical Effectiveness of Tests in Selection: Discussion and Tables." *Journal of Applied Psychology*, 1939, 23, 565-578.

37. Wald, A. *Statistical Decision Functions.* New York: Wiley, 1950.

38. Wiggins, J. S. "Strategic, Method, and Stylistic Variance in the MMPI." *Psychological Bulletin*, 1962, 59, 224-242.

Session II

Theme:
Testing and the Medical Profession

Programmed Testing in the Examinations of the National Board of Medical Examiners

JOHN P. HUBBARD,
National Board of Medical Examiners

Ten years have now passed since the National Board of Medical Examiners came to Educational Testing Service for advice and help in converting our time-honored essay tests to the more modern techniques of multiple-choice testing. The change was accompanied by the cries of those who chose not to understand multiple-choice testing and the criticism of those who, understanding the tests, still did not like them. Nevertheless, with the assistance of those such as John Cowles, then a member of ETS, convincing evidence soon accumulated to demonstrate the gains that had been achieved in the reliability and validity of our written examinations. (1, 2) The new examinations prospered and after relying heavily upon the experience, the facilities, and the excellence of ETS for a period of five years, we were bold enough to strike out on our own. We added to our staff highly qualified individuals from the field of psychometrics and now after another five years, we have welcomed this opportunity to return to ETS at this Conference to describe and—not without some trepidation—to ask for your critical comments about a testing method developed by the National Board.

I have chosen for the title of this presentation, "Programmed Testing." Let me make it clear, however, that I do not wish to become involved in prevalent debates over programmed teaching. Rather this title is intended to suggest that this new testing method has certain features that are similar to those of programmed teaching. Whether one follows Skinner down the linear path, or prefers the branching program method of Crowder, the essential characteristic of programmed teaching, with or without machines, appears to be a step-by-step progression toward carefully constructed goals. Each step calls for specific knowledge. The student must already have the knowledge or he must master it before he may progress to the next step. Similarly, in the testing method that I wish to describe to you, the examinee proceeds in a step-by-step fashion through a sequential unfolding of a series of problems. It is this feature that has, we believe, justified the terminology of our title "Programmed Testing."

Since any test must be viewed in the light of the purpose for which it is designed, let me summarize briefly the objectives of National Board examinations. Our primary objective is to determine the qualification of individual physicians for the practice of medicine. A physician, having successfully completed the extensive series of National Board examinations, may present his certificate to the licensing authority of the state in which he wishes to practice and obtain his license without further examination. If the physician has not elected to take National Board examinations, he must go before the state medical examining board, and if, later in his medical career, he should move to another state, he may be required to repeat this performance perhaps years after he had thought to leave qualifying examinations far behind him. National Board certification is a permanent record and, with few exceptions, permits physicians to move from one state to another without repeated examinations.

This was the initial purpose of the National Board, but it is not its only function. Following the change to multiple-choice testing, and as the reliability, validity, and impartiality of these examinations gained recognition, medical school faculties began to see in these examinations a means of measuring their students class by class and subject by subject, and comparing the performance of their students in considerable detail with the performance of other medical school classes across the country. Thus, these examinations have come to be used widely as extramural evaluations of medical education throughout the United States.

Our examinations are set up in three parts. Part I is a comprehensive two-day examination in the basic sciences usually taken at the time that a medical student is completing his second year of medical school. Part II is a two-day examination in the clinical sciences, designed for the student at the end of the fourth and final year of the medical school curriculum. The third and final part—our Part III—is designed for those who have passed Parts I and II, who have finished their formal medical school courses, have acquired the M. D. degree, and have had some intern experience. It is this Part III examination that is the subject of this presentation today.

Whereas Parts I and II are looked upon as searching tests of knowledge and the candidate's ability to apply his knowledge to the problem in hand, the Part III examination is designed to measure those attributes of the well-trained physician that, rather glibly, we call clinical competence. It has been the long-standing conviction of the National Board that, before we certify an individual to a state licensing board as qualified for the practice of medicine, we should—if we can—test his ability as a responsible physician. Can he obtain pertinent information from a patient? Can he detect and properly interpret abnormal signs and symptoms? Is he then able to arrive at a reasonable diagnosis? Does he show good judgment in the management of patients?

Historically, the National Board sought to answer these questions by means of a practical bedside type of oral examination based upon the candidates' examination of carefully selected patients. In earlier days with few candidates and few examiners, this procedure was effective. More recently, with thousands of candidates, thousands of patients, and thousands of examiners, we found ourselves running into a difficulty that you will be quick to recognize. We were dealing with three variables: the examinee, the patient, and the examiner. Here we had two variables, the patient and the examiner, that we were unable to control at the bedside in order to obtain a reliable measurement of the examinee.

Approximately four years ago, we felt compelled to face up to the necessity of developing a better test of clinical competence or admitting defeat and abandoning the effort. Therefore, we undertook a two-year project with the support of a research grant from the Rockefeller Foundation and the cooperative help of the American Institute of Research. The first step in this project was to obtain a realistic definition of the skills that are involved in clinical competence at the intern level, since it is this level of competence that our Part III examination is intended to measure. The method used for this definition was the critical incident technique under the direct guidance of Dr. John Flanagan. By interviews and by mail questionnaires, senior physicians, junior physicians and hospital residents throughout the country were asked to record clinical situations in which they had personally observed interns doing something that impressed them, on the one hand, as an example of good clinical practice and, on the other hand, an example of conspicuously poor clinical practice. A total of 3,300 such incidents were collected from approximately 600 physicians. This large body of information provided a rich collection of factual information that constituted a profile of the actual experience of interns. We had arrived at a well documented answer to the question of *what to test*? The next step—and a formidable one—was to determine *how* to test the designated skills and behaviors of interns.

Many methods were explored. Motion pictures of carefully selected patients were introduced to eliminate the two variables that had vexed us in the traditional bedside performance. The patient, projected on the screen, became constant and the examiner appeared in the form of pretested multiple-choice questions about the patient. This method has stood up well under the test of usage and continues as a part of the total examination.

A second method—that which we have called Programmed Testing—was evolved to test the intern in a realistic clinical situation where he is called upon to face the unpredictable, dynamic challenge of the sick patient. In real life, the intern may be called to see a patient who, let us say, has just been admitted to the medical ward of the hospital. The intern sees the patient and studies the problem; he obtains information from the patient; he performs a physical examination; and he must then decide upon a course of action. He orders certain laboratory studies and the results of these studies may then lead him to definitive treatment. The patient's condition may improve, or perhaps worsen, or be unaffected by the treatment. The situation has changed; a new problem evolves; and again decisions and actions are called for in the light of new information and altered circumstances.

In the testing method, as we have developed it, a set of some four to six problems related to a given patient simulates this real-life situation in a sequential, programmed pattern. The problems are based upon actual medical records, and may follow the patient's progress for a period of several days, several weeks, or even months until eventually, as in real life, the patient improves and is discharged from the hospital, or possibly has died and ends up on the autopsy table. At each step of the way, the examinee is required to make decisions; he immediately learns the results of his decisions, and with additional information at hand, proceeds to the next problem.

I believe at this point you are detecting certain similarities between the design of this test and the methods of programmed teaching—a step-by-step progression to the goal, each step accompanied by an increment of information upon which the next step depends.

Essential to the methodology of this form of testing, as in the case of programmed teaching, is the concealing of additional information until the examinee has made his decision and has earned the right to have the additional information. We, therefore, first turned to the tab test method. But the tab test, with the tearing off of bits of paper to reveal the underlying information, seemed to us difficult to produce for mass testing and awkward for the examinee.

We also gave serious consideration to the technique for the testing of diagnostic skills described by Rimoldi in a series of papers. (3) Again, a clinical situation is presented and the candidate is offered a number of steps that might be taken to arrive at the correct diagnosis. Each choice appears on a separate card on the back of which is information pertinent to the selected choice. In Rimoldi's hands, the scoring of this test depends not only upon the nature of the choices selected but also the order in which the choices are made. This test, although it has many interesting features, also appeared difficult to handle for mass testing and furthermore does not altogether meet our objectives for a thorough evaluation of diagnostic acumen and judgment in the management of patients.

We then devised the present method, the idea for which has recognizable origins in the tab test and the Rimoldi test but uses a different technique. Instead of tearing off bits of paper or flipping over cards to find the appropriate information, we have concealed the information under an erasable ink overlay. The examinee first studies the problem and a carefully prepared list of possible courses of action. He then makes his decisions and turns to a separate answer booklet where he finds a series of inked blocks each numbered to correspond to the given choices. He removes the ink for selected choices with an ordinary pencil eraser and the results of his decisions are revealed.

At first we had considerable difficulty in finding a printing technique that would permit erasure of the overlying ink block without, at the same time, erasing the underlying information. With the help of an interested specialty printer, a method was developed that has the genius of simplicity. The answers – the results of decisions – are printed and numbered serially in an answer book. A thin acetate layer is laminated on the pages of the answer book and on top of the cellophane layer, blocks of ink are applied to cover the underlying printing. The ink is of a special formula so that when dry it can be removed easily by an eraser or scraper. The acetate interphase layer protects the underlying printing.

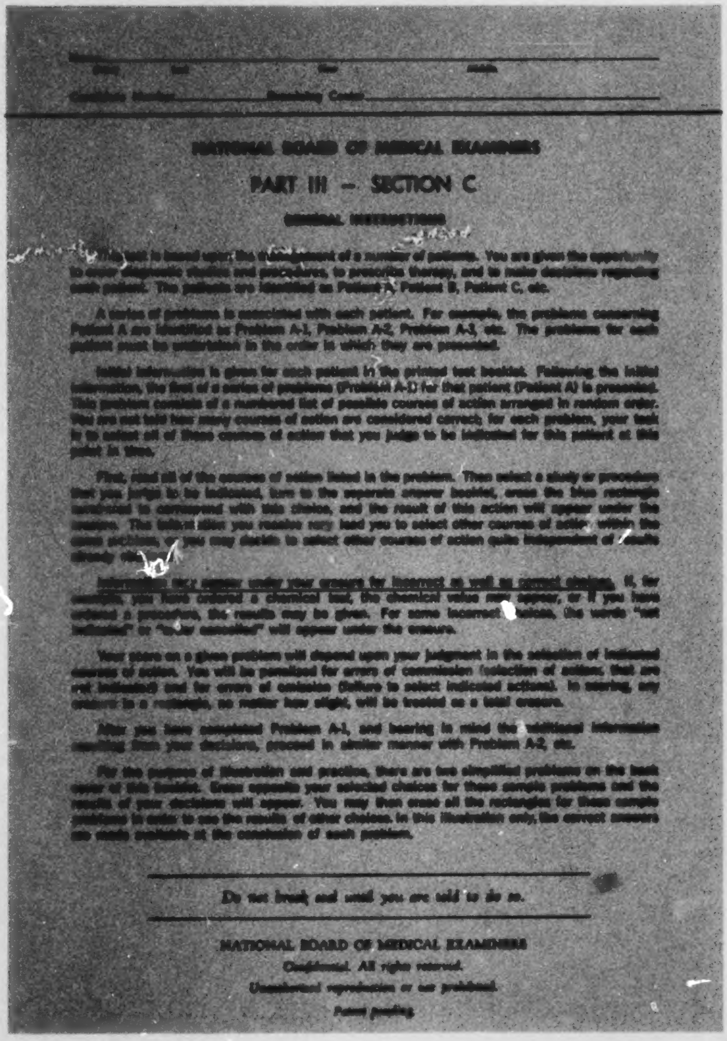
The method is readily adaptable to mass testing and also has the advantage of being foolproof for scoring purposes. The examinee has no way of putting ink back over an answer. If, when he sees the results of his decision, he finds that he has made a wrong choice or if mistaken choices become apparent as the solution to the problem unfolds, he is stuck with the choices he has made. He cannot change his answers and he cannot cheat by peeking ahead under tabs or on the back of cards. His responses, whether right or wrong, are clearly apparent for the scorer to count.

I shall have more to say about scoring, but first a word about content. The complexities of the clinical situations contained in these tests are such as to make them very difficult to describe—especially, I might add, for a non-medical audience. If, however, I may take a leaf from ETS tests, an over-simplified example may be helpful. I have frequently seen on ETS tests an over-simplified example of a multiple-choice item: Chicago is (A) a state, (B) a city, (C) a country, (D) a continent, (E) a village. Just as ETS would, I am sure, resent any implication that this gives a fair impression of the potential of multiple-choice testing, so too the over-simplified example we use for purposes of instruction to the examinee is not to be looked upon as any indication of the difficulty and complexity of the problems in the actual test.

Figure 1, the front cover of one of these tests, is shown to indicate that carefully worded instructions are read aloud by the proctor at the beginning of the test. The candidate is told that the test is based upon his judgment in the management of patients. He is told that initial information is given for each patient and that following the initial information a numbered list of possible courses of action constitutes the first problem for this patient. He is not told how many courses of action are considered correct; his task is to select those courses of action that he judges to be important for the proper management of this patient at this point in time. After he has arrived at a decision on a course of action, he must turn to the separate answer booklet and erase the ink rectangle numbered to correspond to his choice, and the result of his action will appear under the erasure. He is told that information will appear under the erasure for incorrect as well as correct choices. If, for example, he has ordered a diagnostic test, the result of the test will appear under his erasure whether or not the selected test should have been ordered. After having completed the first problem for the first patient he then goes on to the second problem for the same

FIGURE 1

Front Cover of Test Booklet with Detailed Instructions for the Candidate



NATIONAL BOARD OF MEDICAL EXAMINERS

PART III — SECTION C

GENERAL INSTRUCTIONS

Do not break seal until you are told to do so.

NATIONAL BOARD OF MEDICAL EXAMINERS

Confidential. All rights reserved.

Unauthorized reproduction or use prohibited.

Patent pending.

patient and so on throughout the test.

Figure 2 shows the back page of the test booklet. Before breaking the seal of the booklet, he is given a chance to familiarize himself with the method and to practice on two simplified problems related to one patient. At the top of the page is a brief description of a patient who is brought to the emergency room of the hospital in a comatose condition. From the information given, any medical student would recognize the coma as due to diabetes (just as any ETS examinee would recognize that Chicago is a city). The first problem for this patient then offers nine courses of action that call for immediate decision. Three of these nine choices constitute proper management at this point: selection of these three and only these three choices leads to a perfect score for this problem; and, for this sample problem, the key to the perfect score is included on the page in order to give the candidate some feeling of confidence in his understanding of the method.

Figure 3 is an enlargement of choices 3, 4 and 5 in this list. Choice No. 4 is one of the essential procedures. It is shown here with the erasure having been made and the answer revealed. The examinee has decided to catheterize the patient to test the urine and the urine is found to contain large amounts of glucose and acetone, characteristic of the condition with which he is dealing. Let us now assume that he did not recognize diabetes as the cause of the patient's coma and he decided to select choice No. 5 and to perform a lumbar puncture. His erasure for this choice would reveal the words "pressure and cell count normal." Thus, in a very realistic fashion, we are simulating a situation with which an intern might be confronted in the middle of the night when the decisions are entirely his own with no senior physician looking over his shoulder and saying "No, do not do a lumbar puncture." He makes his decision, he performs the action and obtains the results. He may or may not realize as the problem unfolds that the procedure was unnecessary or ill-advised.

Having made his decisions for the immediate steps to be undertaken for this patient, he then proceeds to the second problem. Figure 4 shows an enlargement of Items 10, 11, and 12 appear-

FIGURE 2

Back Cover of Test Booklet with Sample Problems

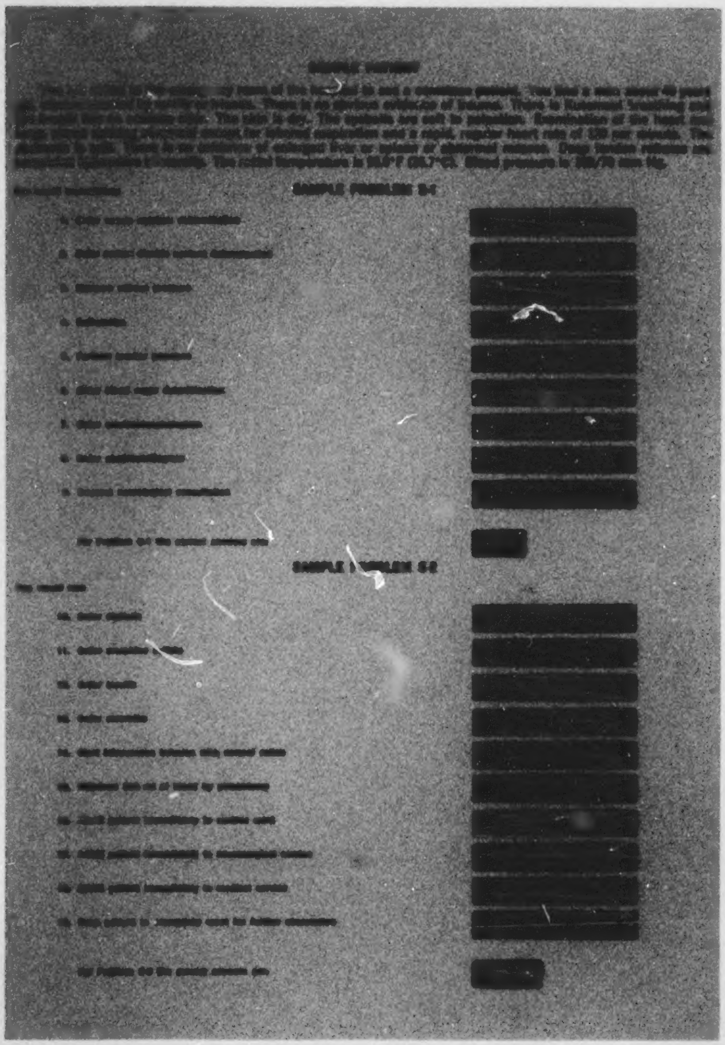


FIGURE 3

Enlargement of Items 3, 4, *and* 5 *of Sample Problem Showing Erasures for Items* 4 *and* 5

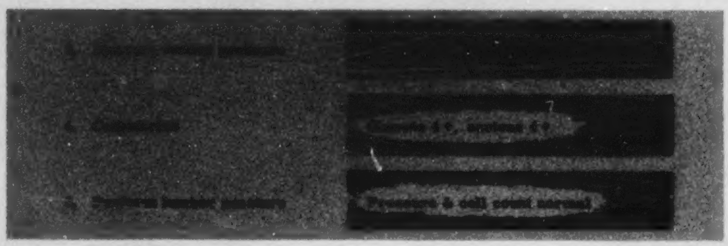
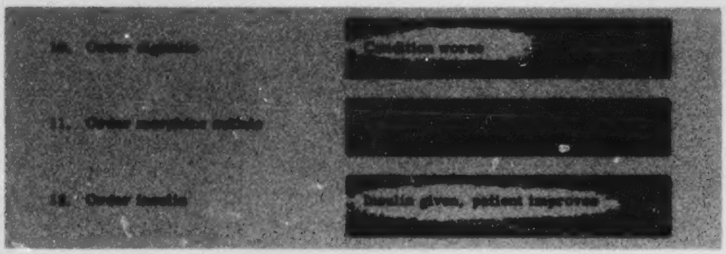

5. Perform lumbar puncture
Pressure & cell count normal

FIGURE 4

Enlargement of Items 10, 11, *and* 12 *of Sample Problem Showing Erasures for Items* 10 *and* 12


10. Order digitalis
11. Order morphine sulfate
12. Order insulin
Condition worse
Insulin given, patient improves

ing in this second problem. Item 12 reads "Order Insulin." This is a correct decision arising from the information that he should have uncovered in the first problem; he erases the corresponding block and sees that the patient's condition improves as a result of his action. But he is also given the opportunity to order other medication as for example, in Item 10, digitalis. This is an incorrect choice that would reflect error in the first problem; if he should order digitalis, he would find, as revealed under his erasure, that digitalis is given in accordance with his orders and the patient's condition worsens. In the actual internship situation, it might not be until the following morning, when the patient is seen by the Chief of Service, that the intern learns of his error in ordering digitalis. The members of our Test Committee, who are physicians prominent in their respective fields and with considerable experience in this manner of testing, sometimes become rather fanciful in the wording of the results of the examinee's actions, particularly with regard to incorrect decisions. One examiner suggested that if the examinee selected a choice that would be considered a fatal error, he should find under the erasure "You have just killed your patient; go on to the next patient."

Now, to turn to the scoring of this test. Let me remind you, as stated earlier, that the basic function of National Board examinations is to serve as a qualification for the general practice of medicine. After having passed the final test of clinical competence in Part III, the candidate is certified and we say in effect: We have examined this individual as carefully as we know how and we consider him qualified to assume responsibility for the medical care of patients. Therefore, although we are interested in excellence, we are mainly concerned with the identification of those few who cannot be considered safe to practice on their own. The focus of this examination is, therefore, on the lower end of the distribution curve.

After having carefully studied several different formulae for the scoring, we arrived at an error scoring to count both sins of omission and sins of commission. Each of the several hundred choices of courses of action offered in the test is classified as to whether it definitely must be done for the well-being of the patient or whether it should definitely not be done and if done would be a serious error in judgment that might be harmful to the patient. A third category includes choices of action that are relatively unimportant, procedures that might be done or might not be done, depending upon local conditions and customs. A candidate who fails to select a choice considered mandatory or who selects a choice considered harmful receives an error score. The choices in the equivocal middle ground receive no score.

Thus, we are dealing with a test and a scoring procedure that are quite different from the usual multiple-choice method in which the examinee is offered a number of choices and directed to select the one best response. Here we offer him a number of choices and require him to use his best judgment in selecting those that he considers important for the management of the patient. Usually, as in a practical situation on the medical ward, he recognizes a number of actions that should definitely be done and other actions that should definitely not be done. His responses are, therefore, interrelated. If he is on the right track, he makes a number of correct decisions among the available choices; then, by his erasures, he gains the information necessary for the proper management of the patient in the next problem and the next set of choices. If he starts off on the wrong track in this programmed test, he may compound his mistakes as he proceeds and he may become increasingly dismayed as he learns from his erasures the error of his ways. But, if he discovers that he is on the wrong track, he has a chance to change his course, although he cannot undo the mistakes he has already made, again a situation rather true to life.

Finally, a brief summary of the statistical analysis of this testing procedure. As you have probably already noted, both the structure of the test and the manner in which we are now scoring it are such as to affect the reliability adversely. In our desire to simulate real-life situations, we have included within each problem a varying number of interrelated responses. Furthermore, there is interdependency between one problem and the next. To return to the two sample problems, anyone who knows anything about diabetic coma would decide to do the three procedures coded as correct for the first problem and he would avoid other procedures coded as incorrect. Then, having confirmed the diagnosis in the first problem, he would have no doubt about the further management of the patient in the second problem. The interdependency of responses within each problem and from one problem to the next has the effect of decreasing the number of points upon which the test score is built and, consequently, decreasing the reliability of the test.

We have studied at some length the balance between the objective to simulate real-life situations in this sequential manner of testing and the objective to obtain high or reasonably high reliability. While the reliability of our tests of Part I and Part II is quite consistently between .80 and .90, the internal consistency measure of reliabiliy for this portion of Part III for the first few administrations has been in the range of .40 to .70 with a mean of .53. We are now taking several steps that may be expected to increase the reliability. The test has been lengthened; the number of items for the next administration in January 1964 has been increased from approximately 200 to approximately 400. The examiners, the experts who construct the tests, are learning from the item analyses the need for more discriminating judgment in categorizing each choice as right, wrong, or equivocal. The task is quite different and considerably more arduous than the more familiar task of deciding on the one best among five choices. On the other hand, the examiners find themselves on somewhat more familiar ground and feel that they are dealing with practical situations in a more realistic manner than when they are faced with the necessity of a single best choice.

We have also looked closely at the correlation between this programmed test of clinical competence in Part III (taken after 6 to 12 months of internship) and the multiple-choice tests of knowledge of clinical medicine in our Part II (taken before internship). The correlation between this portion of Part III and the total Part II was .42 in 1962 and .35 in 1963. Corrected for attenuation, the correlation between these two tests in 1963 would have been .53. These correlations, positive and yet moderate, reflect about the degree of relationship we would expect between medical knowledge and additional elements of clinical competence that are inevitably based upon medical knowledge.

In conclusion, let me summarize by saying that we have developed a testing method that promises to open up new dimensions in evaluating professional competence. We have described this technique as "Programmed Testing" because of features that are similar in principle to "Programmed Teaching," that is to say a step-by-step progression to carefully designed objectives, each step accompanied by an increment of information essential to the sequential unfolding of the problems. The method is far from perfect and needs continuing refinement. It has, however, yielded results that give us reason for increasing confidence in our ability to evaluate effectively certain skills and qualities of professional competence, skills and qualities that we consider essential for certification of a physician's readiness to assume independent responsibility for the practice of medicine.

REFERENCES

1. Cowles, John T. and Hubbard, John P. "Validity and Reliability of the New Objective Tests: A Report from the National Board of Medical Examiners." *Journal of Medical Education*, 29:30-34, June 1954.
2. Hubbard, John P. and Cowles, John T. "A Comparative Study of Student Performance in Medical Schools Using National Board Examinations." *Journal of Medical Education*, 29:27-37, July 1954.
3. Rimoldi, H. J. A. "Rationale and Applications of the Test of Diagnostic Skills." *Journal of Medical Education*, 38:364-368, May 1963.

Alternate Criteria in Medical Education and their Correlates

E. Lowell Kelly,
*Bureau of Psychological Services,
The University
of Michigan*

Introduction

The concern of the medical profession with testing and the more general problem of assessment is three-fold:

A. in the evaluation of applicants seeking a license to practice medicine in a state;

B. in the evaluation of applicants to medical schools;

C. in the evaluation of the outcomes of medical education.

The right to practice medicine is a privilege controlled by licensure in each of the states and territories. Although requirements vary widely from state to state, practically all require the completion of a specified program of medical education leading to some type of doctoral degree (in most cases the M. D.) *and* satisfactory performance on an examination designed to evaluate medical knowledge and competence. Such examinations ordinarily consist of a number of parts and are collectively called "State Boards." Since these examinations are typically constructed and graded locally, it is not surprising that their nature varies widely from state to state. Furthermore, because of the jealousy of the states in guarding the right to grant licenses to practice in each state, any physician moving from one state to another, or one who wishes to practice in two adjacent states simultaneously, finds it necessary to submit to two or more unique sets of examinations. While a few states have entered into reciprocal agreements, i.e.—agreed to recognize the validity of the other's licensing examination, such reciprocation is sufficiently rare as to have led to the development of the program of National Board examinations.

The need for, and use of, tests in the evaluation of applicants to medical schools is of more recent origin. Those of you who are familiar with the history of the medical profession will remember that until relatively recently, medicine, like law, was an art acquired by apprenticing oneself to an older and more experienced member of the profession, reading a few books, and passing the state licensing examination. There were a few medical schools associated with certain of our older universities, but only a relatively small proportion of the practicing physicians of the day ever attended them. During the nineteenth century there was a very rapid growth in the number of colleges and schools offering professional training in medicine, but a large proportion of them were proprietary, with the result that their owners were more interested in attracting students for the tuition which they would bring with them than for their aptitude for the study of medicine. This state of affairs is reflected by the fact that in 1904 there were twice as many medical schools in the United States as there are in 1964! As a result of the survey of medical education sponsored by the Carnegie Foundation, which culminated in the famous Flexner report (1), two very significant changes occurred in professional training for medicine: (a) the standards of medical education were markedly increased; and (b) the medical schools began more and more to seek an affiliation with a university. Today there are only a few non-affiliated schools remaining.

Although medical practicioners had always been accorded a fairly high status in their communities, these developments served to enhance further the prestige associated with medicine as a profession. As a result, membership in the profession became the aspiration of many more younger people than could be accommodated in recognized medical schools. Thus, the faculties of these institutions found themselves confronted with the necessity of selecting the most promising applicants for the study of medicine. It is not surprising, therefore, that medical schools were among the first to utilize a professional aptitude test. The present Medical College Aptitude Test, which was developed in 1927, was an outgrowth of the Moss Medical Aptitude Test and the Professional Aptitude Test. (2) Since 1935 it has been given every year in several hundred pre-medical schools to at least 90 per cent of all applicants for admission to medical schools throughout the United States.

The affiliation of medical schools with universities and, even more important, the introduction of extensive components of basic science teaching in the medical curriculum led to an increasing concern on the part of medical school faculties with the evaluation of student achievement and assigning grades. There is fully as much disagreement among medical school professors as among teachers everywhere about the best methods of evaluating students and assigning grades. In fact, it was inevitable in these professional schools, where grades are so important in determining not only survival but assignment to internships and other professional opportunities, that there would be much ferment and discussion regarding the comparative values of oral versus written tests, of objective versus essay tests, of tests emphasizing short-term versus long-term learning, whether grades should represent progress made as a result of taking a course or the absolute level of accomplishment upon course completion, of whether grades should be given primarily for factual learning or for the demonstration of professional skills, and so on.

While the ratio of the number of applicants to the number of available places in the admission class has declined over the last few years, medical schools are generally more concerned with the selection of their students than are universities and colleges or even other professional schools. This is true for two very good reasons: First, the high cost of constructing, equipping, and staffing medical schools results in a very high per-student societal investment as compared with other types of educational institutions. Second, because of the integral nature of the curriculum in medicine, it is generally not feasible for medical schools to admit students in the second, third, and fourth years. Thus, any beginning student who does not succeed in the program leaves a gaping and costly hole in an establishment geared for a certain number of students. This, in turn, subjects the institution to public criticism for not turning out as many doctors as it was tooled up to do. The typical college or university may regret the loss of an entering freshman but it can always fill its upper classes with transfer students and thus utilize to the fullest the educational resources of the institution. The result is that neither the institution nor society is so painfully aware of the importance of good selection of beginning students as are medical schools.

Still another unique situation has contributed to the extensive concern of medical schools with testing. To a degree that is not true of any other category of professional education, the Association of American Medical Colleges monitors and coordinates the typical multiple applications of medical school candidates and provides feedback concerning the quality of students entering each medical school. This has served to sensitize the admissions committees of medical schools to very wide differences in both the quantity and quality of applicants to different schools. The result is that medical schools in general probably invest more time and money in the evaluation of applicants than any other educational institution. For example, most medical schools have fairly large admissions committees whose members are responsible for interviewing all applicants to the school, (3, 4) and make an eventual decision to admit or not admit an applicant only after an extended staff conference regarding each applicant.

These, then, are the factors that have combined to develop an increasing concern on the part of the medical profession with the problems of testing.

My personal involvement in the problem of selecting medical students began about a dozen years ago, just about the time that Fiske and I completed our project on the selection of graduate students in clinical psychology. (5) I was approached by the late Wayne Whittaker, assistant dean of the University of Michigan Medical School and Chairman of the Admissions Committee, who asked me to work with him to improve the selection of students for our medical school. Because I found that he and his committee were deeply concerned with selecting not only students who would succeed academically, but also those who would become good physicians willing to accept social responsibilities commensurate with society's investment in them, I was delighted to accept his invitation to participate in a collaborative study.* With a small research grant we carried out a preliminary study of the senior class of 1952 and planned a more intensive study of students entering in the fall of 1952—the class that graduated in 1956. The findings, which I am reporting here, are based, for the most part, on 112 of the 181 members of the class of 1956. The fact that this group does not represent the entire class was primarily a function of class schedules rather than any biased selection of the sample.

Our broad objective was simply to try to improve the over-all quality of the students selected to receive medical training. In the hope of identifying variables that should be considered at the time of the selection decision, we made an intensive study of the "mistakes" of the admissions committee, i.e.—those students who had been admitted but failed in the course of their training. We eventually accumulated data regarding 100 potential predictor variables.

For our criteria, we began, of course, with the most convenient and frequently used index of academic performance, the grade point average. Medical educators, like others, are free to admit that academic grades are not the only, and perhaps not the most important, criterion of success in medical education. Nevertheless, grades are regarded as important by both students and staff, and successful academic performance, especially in the first two years (pre-clinical), is a *sine qua non* for developing one's clinical skills in the later years of medical training. Therefore, as soon as the first-year grade point average had become available for the class, correlational analyses were begun.*

---

* I have purposely postponed publication of certain ofthe findings until changes in staff, curriculum, and grading practices will make it impossible to point an accusing finger at any specific department or faculty member!

Twenty-three of the 100 predictor variables were found to be significantly correlated ($p < .05$) with the first year GPA. Two variables, All Pre-med. Grades and Pre-med. Science Grades, tied for first place as the best predictor of this criterion, both yielding an r of +.61. This variable was also significantly predicted by the four subscores of the MCAT with coefficients ranging from .25 to .30. Offhand, this would seem to reflect a relatively satisfactory state of affairs. However, members of the admissions committee were not satisfied. Even with validity coefficients of this magnitude considerable error of prediction remains, and, as noted above, the desire not to lose already admitted students is very strong. Of greater concern to members of the admissions committee, however, was the finding that their individual ratings of the applicants yielded validity coefficients lower than that provided by a simple average of all pre-med. grades. The actual coefficients of five members ranged from +.27 to +.59 in spite of the fact that the ratings were made by persons who had had an opportunity to study the entire pre-med. transcript, review the profile of MCAT scores, read the letter of recommendation, and discuss each case at a staff conference at which the interview impressions of at least one of the committee members were reported.

Quite obviously, something was wrong. There were two possibilities: (a) members of the admissions committee were not identifying and/or properly weighting relevant items from the large mass of information available to them; or (b) the criterion of first-year grades was not the appropriate one against which to check the validity of their individual judgments. I am sure that you can guess which of these alternatives was chosen by members of the admissions committee. While they were, of course, concerned with evaluating the aptitude of the student to successfully complete the prescribed course work of the medical curriculum (i.e.—make grades), they were much more concerned, they insisted, with selecting those applicants who also had the other characteristics essential to becoming a good physician. It would, therefore, be necessary to secure additional and very different criterion measures of success in medicine before the unique validities of the ratings derived from this elaborate admissions procedure could be properly evaluated.

---

* The writer gratefully acknowledges the assistance of Gordon Bechtel, Lillian Kelly, and Leonard Uhr who served as research assistants at various stages of the study.

During the next three years much time and effort were devoted to the development and acquisition of alternative measures of achievement and performance in medicine. Eventually data became available for 54 criteria.

With 100 predictor variables and 54 criterion variables, several alternate modes of analyses suggested themselves and, thanks to the availability of the high speed computer, several of them were carried out. In this paper we shall be primarily concerned with an analysis of the resulting 5400 correlations between the 100 predictors and the 54 criteria. More specifically, we shall concern ourselves with the relative utility of the predictors, i.e.—the number of alternate criteria which they predict, and with the significant correlates of each of the 54 criteria. I have intentionally selected the term *correlates* of the criteria rather than *predictors* because of the obvious limitations of the present study with respect to generalizability. With such a large number of variables and an N of only 112, it would have been possible to have computed spuriously high multiple correlations to predict many of the criteria, correlations that would certainly have shrunk markedly if the resulting regression equations had been applied to another class. Furthermore, it must be remembered that I am reporting data not only for a single class but also for but one medical school. In view of the known differences not only in the quantity and quality of applicants but in selection procedures used by different schools, any attempt to make generalizations regarding predictive value of specific variables for other institutions would be extremely hazardous. In spite of these limitations, I believe that our findings are worthy of serious attention, not because of their immediate applicability to the problem of selection, but rather because of their implications for the problems of testing and measurement in all educational institutions, of which the medical school is but one example.

Potential Predictor Variables

Table I lists the 100 potential predictor variables selected for analysis; also shown is the number of the 54 criteria with which each variable showed a correlation of at least .20, i.e.—a value yielding a P of $< .05$. Since there were 54 possible significant correlations for each predictor variable, chance alone would yield an expectancy of two to three such correlations for each predictor.

Part A of Table I lists 12 predictor variables which have been labelled Intellectual and Cognitive. This category includes pre-med. grades, MCAT scores, and ratings by the five individual members of the admissions committee, since these ratings appear to be primarily determined by the pre-medical academic record. Similarly, the month of acceptance is included in this category of variables because of the practice of according early admission to the applicants rated most favorably by the admissions committee. In general, it will be noted that these intellectual and cognitive variables yield significant correlations with a fairly large proportion of the 54 criterion variables.

Part B of Table I lists the 88 non-cognitive predictor variables used. The first subgroup of these includes 19 background variables derived from analysis of the application blank or the transcript of pre-medical college training. As will be noted, most of this group of variables predict more than a chance number of the 54 criteria. Note that the number of credit hours in biology appears to be the best of these predictors, yielding significant correlations with 12 of the 54 criteria. There follows an extensive list of measures of personality characteristics and interests. The first block includes the 16 scores derived from the Cattell 16 Personality Questionnaire Test. (6) The labels given these factors correspond to the positive or high scoring end of each scale. Incidentally, these Cattell scores were based on an administration of the test to the class as seniors, whereas

Table I

100 Potential Predictor Variables of Performance in Medicine and the Number of 54 Alternative Criteria with which Each was Significantly Correlated

A. Intellectual and Cognitive Variables

		No. of Criteria
1.	All Pre-Med Grades	28
2.	Pre-Med Science Grades	27
3.	Rating: Adm. Comm. Member No. 1	26
4.	Rating: Adm. Comm. Member No. 2	30
5.	Rating: Adm. Comm. Member No. 3	23
6.	Rating: Adm. Comm. Member No. 4	20
7.	Rating: Adm. Comm. Member No. 5	9
8.	Month Accepted	17
9.	MCAT: Verbal	17
10.	MCAT: Quantitative	9
11.	MCAT: Modern Society	15
12.	MCAT: Science	14

B. Non-Cognitive Variables
Background Variables

1.	Year of Birth	8
2.	Own a Car	4
3.	Father's Occupation	8
4.	Father's Education	8
5.	Mother's Education	4
6.	Percent Self-supporting	5
7.	Reported Est. of Summer Earnings	8
8.	Reported Est. of Religious Activity	3
9.	Am't of Pre-Med (3 or 4 yrs)	5
10.	Marital Status	2
11.	Month Applic. Submitted	4
12.	Height	3
13.	Weight	3
14.	No. of Credit Hrs. English	3
15.	No. of Credit Hrs. Foreign Language	2
16.	No. of Credit Hrs. Inorganic Chem.	8
17.	No. of Credit Hrs. Organic Chem.	4
18.	No. of Credit Hrs. Physics	3
19.	No. of Credit Hrs. Biology	12

Cattell Personality Factor Questionnaire Variables

1.	Cattell 16 P-F: Cyclothymia	8
2.	Cattell 16 P-F: Gen. Intell.	4
3.	Cattell 16 P-F: Ego Strength	4
4.	Cattell 16 P-F: Dominance	17
5.	Cattell 16 P-F: Surgency	5
6.	Cattell 16 P-F: Super-Ego	1
7.	Cattell 16 P-F: Adventurous Cyclothymia	12
8.	Cattell 16 P-F: Emot. Sensitivity	5
9.	Cattell 16 P-F: Paranoid	8
10.	Cattell 16 P-F: Hysterical Unconcern	1
11.	Cattell 16 P-F: Sophistication	11
12.	Cattell 16 P-F: Anx. Insecurity	4
13.	Cattell 16 P-F: Radicalism	10
14.	Cattell 16 P-F: Indep. Self-Sufficiency	12
15.	Cattell 16 P-F: Will Cont. & Stability	1
16.	Cattell 16 P-F: Nervous Tension	10

Strong VIB Variables

		No. of Criteria
1.	Artist	7
2.	Psychologist	9
3.	Architect	8
4.	Physician	2
5.	Osteopath	5
6.	Dentist	4
7.	Veterinarian	10
8.	Mathematician	7
9.	Physicist	7
10.	Engineer	5
11.	Chemist	10
12.	Production Manager	2
13.	Farmer	7
14.	Aviator	6
15.	Carpenter	6
16.	Printer	12
17.	Math. Phys. Sci. Teacher	11
18.	Ind. Arts Teacher	8
19.	Voc. Agric. Teacher	7
20.	Policeman	10
21.	Forest Service Man.	4
22.	YMCA Physical Dir.	4
23.	Personnel Dir.	0
24.	Public Admin.	2
25.	YMCA Secretary	1
26.	H. S. Soc. Sci. Teacher	6
27.	City School Supt.	1
28.	Minister	4
29.	Musician	2
30.	C. P. A.	10
31.	Senior C. P. A.	4
32.	Accountant	1
33.	Office Mgr.	2
34.	Purchasing Agent	2
35.	Banker	8
36.	Mortician	10
37.	Pharmacist	6
38.	Sales Manager	9
39.	Real Estate Salesman	7
40.	Life Insurance Salesman	6
41.	Advertising Man	7
42.	Lawyer	5
43.	Author-Journalist	4
44.	Pres. Mfg. Concern	6
45.	Interest Maturity	3
46.	Occupational Level	2
47.	Masculinity-Femininity	1
48.	"Anxiety"	9
49.	"Theoretical Values"	8
50.	"Economic Values"	4
51.	"Self-Confidence"	2
52.	"Sociability"	5
	Also McQuitty Health Index	7

the 52 scores derived from the Strong Vocational Interest Blank (7) and all other variables of Table I were based on instruments administered before admission to medical school, i.e.—under conditions which led applicants to perceive them as a part of the total process of admissions.

The first 47 variables derived from the Strong VIB are the familiar vocational interest scores which were coded on the basis of numerical scores rather than letter grades. In general, high scores reflect a pattern of interests highly congruent with those of persons successfully engaged in each profession or occupation. Variables 48 to 52 were also derived from responses to the Strong VIB using empirically derived scoring keys to assess personality variables alternatively measured by the Taylor Manifest Anxiety Scale of the MPI, (8) the Allport-Vernon Scale of Values, and the Bernreuter Personality Inventory. (9) Finally, the McQuitty Health Index (10) was based on responses to a self-report form developed to assess personality integration.

Although none of these non-cognitive variables correlated significantly with a very large number of the 54 predictors, it will be noted that most of them yield more than a chance number of significant correlations. Thus, nine or more of the 54 criteria were found to be significantly correlated with the Cattell factors of Dominance, Adventurous Cyclothymia, Sophistication, Radicalism, Independent Self-Sufficiency, and Nervous Tension. A fairly impressive number of significant correlations also appeared for the interest patterns of Psychologist, Veterinarian, Chemist, Printer, Math-Physical Science Teacher, Policeman, CPA, Mortician, and Sales Manager. Finally, the "Anxiety" score derived from the Strong yielded nine significant correlations with criterion measures.

The Criteria and their Correlates

Table II summarizes the 54 criteria used in this study. Shown also for each criterion is the number of the potential cognitive and non-cognitive predictor variables with which it was significantly correlated. Column 3 indicates the cognitive variable yielding the highest, and column 4 the non-cognitive variable

Table 2.

The Correlates of 54 Alternative Criteria of Performance in Medicine

Criteria	No. of Significant r's with: 12 Cognitive Variables	88 Non-Cognitive Variables	Highest Correlated Variables: Cognitive		Non-Cognitive	
A. Grades						
1. 1st-yr. GPA	12	11	Pre-Med Grades	.61	Veterinarian	.24
2. 2nd-yr. GPA	12	8	Pre-Med Grades	.56	"Anxiety"	.30
3. 3rd/yr. GPA	11	4	Pre-Med Grades	.47	"Anxiety"	.31
4. 4th-yr. GPA	6	5	Pre-Med Grades	.31	No. of Hrs. Biology	-.22
					Dominance	-.22
5. Over-all GPA	12	8	Pre-Med Grades	.57	"Anxiety"	.28
6. Pub. Health (2nd yr)	5	7	Pre-Med Grades	.45	Pharmacist	.26
7. Medicine (4th yr)	5	1	Month Accepted	-.33	No. of Hrs. Foreign Language	.22
8. Surgery (4th yr)	5	2	MCAT (Quant.)	-.20	Dominance	-.36
9. Ob. & Gyn. (4th yr)	0	4	[Cattell "Intell."	.13]	Theor. Values	-.25
					Self-Sufficiency	-.25
10. Pediatrics (4th yr)	7	6	Pre-Med Grades	.38	No. of Hrs. Biology	-.31
11. Psychiatry (4th yr)	0	12	[MCAT (Verbal)	.14]	Mother's Educ.	.28
B. Nationally Administered Tests						
1. Cancer Exam.	11	18	MCAT (Mod. Soc.)	.45	"Anxiety"	.44
National Boards						
2. Medicine	7	12	MCAT (Science)	.37	Insecurity	.24
3. Surgery	12	29	MCAT (Science)	.40	Adventurous	
			Pre-Med Grades	.40	Cyclothymia	-.38
4. Ob. and Gyn.	10	6	Pre-Med Grades	.39	Pub. Admin.	.28
5. Public Health	10	5	MCAT (Science)	.42	Farmer	.22
6. Pediatrics	10	10	MCAT (Science)	.40	Banker	-.27
7. Nat. Bds. Total	10	25	MCAT (Science)	.48	Printer	.32
C. State Boards						
1. Anatomy	8	4	Month Accepted	-.38	Adv. Cyclothymia	-.24
2. Hist. and Embry.	6	8	Member No. 4	.30	Lawyer	.28
3. Physiology	0	8	MCAT (Verbal)	.14	Own Car	-.25
4. Chem. & Toxicology	4	7	MCAT (Verbal)	-.21	Physician	-.31
5. Bacteriology	2	30	Member No. 5	.29	Sales Mgr.	-.38
6. Pathology	0	8	Cattell "Intell."	-.30	Dominance	-.32
7. Hygiene	7	5	Member No. 5	.34	Hrs. Org. Chem.	-.27
8. Practice	0	8	MCAT (Mod. Soc.)	.17	Theor. Values	-.24
9. Med. Jurisprudence	0	12	Gen. Intell.	-.26	Occup.	.30
10. E. E. N. T.	0	0	MCAT (Science)	.13	Occ. Level	.19
11. Obstetrics	1	1	MCAT (Verbal)	-.21	Weight	-.21
12. Surgery	8	3	Pre-Med. Sci.	.24	Surgency	-.31
13. Gynecology	0	5	MCAT (Verbal)	-.14	No. Hrs. Physics	.27
14. Materia Medica	0	1	MCAT (Med. Soc.)	.19	Further Educ.	.20

	No. of Significant r's with:		Highest Correlated Variables			
Criteria	12 Cognitive Variables	88 Non-Cognitive Variables	Cognitive		Non-Cognitive	
D. Sociometric Choices as Seniors						
1. Camping Companion	1	6	MCAT (Science)	-.23	Self-Sufficiency	-.33
2. Office Partner	3	17	Pre-Med Grades	.28	Chemist	-.29
3. Research Promise	8	10	Pre-Med Grades	.41	Dominance	-.27
4. Intimate Friend	0	10	MCAT (Verbal)	-.12	Mortician	.31
5. Phys. to own Family	6	9	Pre-Med Sci. Gr.	.39	Dominance	-.31
6. Colleague Hosp. Staff	4	8	Pre-Med Grades	.30	Dominance	-.26
7. Hosp. Teaching Staff	7	4	Pre-Med Sci. Gr.	.40	Self-Sufficiency	-.23
8. Highest Income	1	21	MCAT (Quant.)	-.21	Sales Mgr.	.37
9. Pers. Satis. as G. P.	2	20	MCAT (Verbal)	-.29	Father's Educ.	-.47
10. Med. Sch. Teacher	8	0	Pre-Med Grades	.35	Nervous Tension	.23
11. Int. in Public Health	4	15	Pre-Med Grades	.32	Dominance	-.32
12. Willing to Accept Salaried Position	0	16	MCAT (Quant.)	.14	Height	-.31
13. Disease (i. e. Specialty) Orientation	3	13	Member No. 2	.31	Father's Educ.	.47
14. Hosp. Administrator	4	7	Pre-Med Grades	.32	Radicalism	-.23
E. Internship Ratings						
1. Personal Appearance	1	1	MCAT (Med. Soc.)	-.22	Am't Rel. Act.	.23
2. Desire to Learn	5	17	Pre-Med Sci. Gr.	.32	Math. Phys. Sci. Teacher	.36
3. Over-all Med. Knowledge	4	7	Pre-Med Grades	.23	Dominance	-.32
4. Diagnostic Competence	0	2	[Pre-Med Grades	.19]	Math. Phys. Sci. Teacher	.32
5. Integrity	1	11	[MCAT (Quant.)	-.20]	Math. Phys. Sci. Teacher	.25
6. Sensitivity to Patients' Needs	0	10	[MCAT (Med. Soc.)	-.18]	Dominance	-.30
7. Abil. to Inspire Confidence	0	2	[Pre-Med Grades	.11]	Dominance	-.26
8. Over-all Promise	0	0	[Pre-Med Grades	.11]	Sales Mgr.	.17

yielding the highest correlation with each of the criteria.

These 54 criteria have been grouped into five categories, A through E. Those in A require little further description than is provided by their names. The grade in the second-year course in Public Health was selected as the criterion because this course at the time was generally regarded by the students as both difficult and not very relevant to their training. The several fourth-year course grades presumably reflect the faculty's best evaluation of the performance of the medical student in the clinical, as contrasted with the pre-clinical, years of medicine. As far as could be determined, these grades were based not so much on tests as on the impressions made by the student as a participant in ward rounds, conferences, and seminars.

As will be noted, the best predictor of medical school grades throughout the four years is the Pre-med. Grades (Average). Of the non-cognitive correlates, the "Anxiety" score derived from the Strong VIB appears most frequently. Whereas there is a relatively high intercorrelation (about .80) between the grade point averages for the first three years of medical school, the situation for the fourth-year course grades is quite different. First-year and fourth-year grades correlate only .53 and the median intercorrelation among the six fourth-year grades is only .22. It is therefore not surprising that a very different pattern of correlates emerges for these fourth-year course grade criteria. As will be noted in several instances, a non-cognitive variable is more closely associated with grades in these courses than a cognitive variable, suggesting the degree to which these grades are assigned on the basis of impressions made by the students while on a particular service and thus are more a function of the student's personality characteristics than of his intellectual performance.

Category B of the criteria includes scores made by the students on nationally administered objective tests. In general it will be noted that performance on these objective tests at the end of medical training is predicted by most of the cognitive predictor variables, best by the MCAT Science score and Pre-med. Grades. It is of considerable interest, however, that grades on these objective tests are also significantly correlated with several non-cognitive predictor variables. For example, 29 of the 88 non-cognitive variables are significantly associated with National Board scores in Surgery, the correlation with one of them being almost as high as that of any cognitive variable.

We now turn to a consideration of the criteria listed in Part C of Table II, marks on the State Board examination (State Boards), required for licensure in Michigan. As contrasted with the National Boards, which are objective examinations, these are typically essay examinations prepared by experienced and often older physicians who volunteer to prepare and grade examinations in each of the 14 subject matter areas. Whereas the median intercorrelation among the scores on the National Boards is +.51, the modal intercorrelation among the scores on the 14 parts of the State Board examination is zero; the median is only .10 and 16 of the intercorrelations are negative! This being the case, it is not surprising to find markedly different patterns of correlates of the grades on the various subparts of the State Boards. For seven of the 14, we note no significantly correlated cognitive predictor variable. And in general for each a non-cognitive variable correlates about as highly as a cognitive variable, suggesting that even though these are written examinations, marks are determined in part by the student's personality characteristics: interest, values, and background variables.

Part D of Table II lists 14 variables derived from Sociometric Choices made by members of the class of 1956 near the completion of their medical training. In our search for more relevant criteria (and, hopefully, for criteria more predictable from the ratings of the medical admissions committee!) we decided to capitalize on the rather extensive opportunities which medical students have to become acquainted with each other's strengths, weaknesses and special competences. In brief, these sociometric ratings were collected as follows: all members of the senior class were assembled in one room, provided with a list of all members of their class, and before they knew what was to follow, they were asked to star the names of the 40 fellow class members whom they felt they knew best. They were next asked to select the three most desirable (or most likely) and the three least desirable (or least likely) persons out of this group of 40 fitting each of the categories indicated by the labels associated with these 14 criteria (Cf. Table II, Part D). The score for each student on each criterion was simply the algebraic sum of the number of positive and negative choices on each item.

It is of interest to note that practically all of these sociometric criteria are significantly associated with far more than a chance number of the potential predictor variables. In fact, most of them can be predicted about as well as any of the categories of criteria.

Furthermore, the pattern of the significant correlates seems to make sense in that those sociometric criteria most obviously associated with intellectual performance are most likely to be correlated with cognitive variables, whereas those primarily related to social acceptability are more often correlated with non-cognitive variables. Finally, the pattern of the correlates makes sufficiently good sense to suggest that these sociometrically derived criteria may have considerable validity for real-life performance.

Our final effort to secure additional and still more relevant criteria of performance as a physician is reflected in the Internship Ratings listed in Part E of Table II. With the assistance of a number of members of the medical school staff, these eight variables were selected as those believed to be most relevant and the most ratable by the supervisors of medical school graduates during their year of internship. We note immediately that these criterion measures tend to be less often significantly correlated with the predictor variables than was the case for sociometric criteria. Only two of them, rated "Desire to Learn" and "Over-all Medical Knowledge" have more than a chance number of correlates among the cognitive predictors. Although each of them tends to be more closely associated with some non-cognitive predictor than with a cognitive one, the general magnitude of the correlations tends to be low. Finally, we note that for the most global of these ratings, "Over-all Promise," there are no significant correlates whatsoever. Apparently this rating, made by many different supervisors in different internship setttings, reflects such a composite of unsystematically weighted values as to result in it not being significantly related to any of the 100 predictor variables.

In summary, most of the criteria were found to have a number of correlates among predictor variables available before admission to medical school. In fact, most of the criteria could be reasonably well predicted by the weighted combination of some subset of predictor variables. Unfortunately, however, because of the relatively low intercorrelations among the alternative criteria, a different set of predictor variables would be needed to select applicants likely to rank high on alternate criteria. We have already noted the extremely low intercorrelations among parts of the State Board examination. The problem of the validity of alternate criteria is even more dramatically demonstrated by examination of the intercorrelations of presumably alternative criteria of the same type of accomplishment. For example, the fourth-year course grade correlations with National Board scores are as follows: Pediatrics .37, Medicine .33, Surgery .19, Obstetrics and Gynecology .12. And, as might be expected, neither of these criteria measures correlates significantly with State Board examinations bearing the same label! Under the circumstances, it is somewhat surprising that most of these criteria are at all predictable from data obtained before admission to medical school.

The Criterion Correlates of the Most Promising Predictor Variables

From Table I, we noted that those variables listed in category A, Intellectual and Cognitive, are generally the most promising predictors of a large number of criterion variables. In fact, all 12 of them yield significant correlations with nine or more of the 54 criterion variables, most typically appearing as the best predictor of the more intellectually loaded criterion measures. The most promising intellectual predictor for this particular group of students was the average of all pre-medical grades, rather than the average of the pre-medical science grades only, as members of the admissions committee had anticipated. In general, the ratings of the members of the admissions committee, here categorized as intellectual or cognitive variables, showed significant correlations with a fairly large number of criterion variables but most typically with those which might be categorized as reflecting intellectual or academic accomplishment rather than those reflecting performance as a physician. Only rarely did these ratings of individual committee members turn out to be as predictive of any criterion as one or more of the pieces of information available to the person making the rating!

The most likely explanation of this attenuated potential validity of clinical judgments of academic performance appears to be a function of the background variable B 19 of Table I. The number of credit hours in biology is the only one of these background variables associated with as many as nine criterion variables. Since the medical school at that time required all applicants to present 12 credits of biology, this variable represented the extent to which applicants presented credit hours in biology in excess of this minimal requirement. Many pre-med. students, especially if their over-all academic record is not good, are encouraged to take additional credits in biology as an indication of their strong interest in medicine and because they are of the opinion that members of the admissions committee would be favorably impressed with a transcript reflecting elected courses in biology. This turns out to have been the case. In general, the ratings of the members of the admissions committee tend to be positively correlated with the number of hours of biology. However, this same variable, number of hours in biology, yielded a significantly negative correlation with 12 of the 54 criteria! These correlations were as follows: State Board Hygiene -.22; 2nd-yr. GPA -.27; 3rd-yr. GPA -.24; 4th-yr. GPA -.22; over-all GPA -.23; grade in Public Health, second-year -.20; fourth-year grade in pediatrics -.31; National Cancer Exam. -.23; National Boards of Medicine -.23; National Boards Public Health -.20; National Board Pediatrics -.22; National Board Over-all -.21. In a word, the potential validity of the clinical prediction of academic success was seriously attenuated by the fact that the members of the admissions committee were noting the number of hours of biology as a relevant predictor but weighting it positively rather than negatively!

Of the non-cognitive variables, the one yielding the largest number of significant correlations with the criteria used was Cattall's factor labelled Dominance-Ascendance versus Submissiveness. These correlations were as follows: State Board Physiology -.23; State Board Pathology -.32; Office Partner -.24; Worthy Recipient of Research Grant -.27; Intimate Friend -.21; Physician to own Family -.31; Colleague-Hospital Staff -.26; Hospital Teaching Staff -.20; Personal Satisfaction as GP -.37; Interest in Public Health -.32; Hospital Administrator -.23; 4th-yr. GPA -.22; 4th-yr. grade in Surgery -.36; Desire to Learn -.27; Over-all Medical Knowledge -.32; Sensitivity to Patients' Needs -.30; Ability to Inspire Confidence -.26. Since a high score on this variable reflects a tendency to be self-assertive, boastful, conceited, aggressive and pugnacious, it appears that those students characterized by submissiveness, modesty, and complacency are more likely to be positively evaluated on these generally noncognitive criterion variables. Another personality variable measured by the Cattell 16 PR shows a similar pattern of negative correlations with 12 of the criterion measures. It is Adventurous Cyclothymia representing a continuum characterized by adventurous versus shy, timid; gregarious versus aloof; and frank versus secretive. In general, withdrawn cyclothymia seems to be more highly prized in this particular subculture, the only significant positive correlation being .29 with the sociometric choice, Likely to Make the Highest Income.

Four additional scores from the 16 PF yielded significant correlations with 10 or more of the criterion variables. These were: Sophistication (all positive correlations except with State Board Gynecology); Radicalism (generally negative correlations except with National Board Over-all); Independent-Self-Sufficiency (generally positive correlations with intellectually loaded criteria and negative ones with sociometric choices involving interpersonal relations), and Nervous Tension (generally positive correlated with intellectually loaded criteria).

Of the 52 variables derived from the Strong, seven were found to be significantly correlated with nine or more of the criteria. Psychologist scores are typically negatively correlated with sociometric choices; Veterinarian scores are positively correlated with ten criteria, including Physician to own Family; Chemist scores are negatively correlated with several sociometric choices but positively with Disease or Specialty Orientation, Willingness to Accept Salaried Job, and National Board scores; Policeman scores are positively correlated with 10 criteria, mostly sociometric choices and intern ratings. CPA scores typically yield negative correlations with criteria. Sales Manager scores are negatively correlated with State Board in Bacteriology, Interest in Public Health, Willingness to Accept a Salaried Job, with National Boards in Surgery and National Boards Over-all; Sales Manager scores are positively associated with State Boards in Practice, Likely to Make Highest Income, 4th-yr. Grade in Psychiatry and Sensitivity to Patient Needs as rated by the Intern Supervisor.

Another Strong VIB score, Printer, proved to be a relatively good predictor of several different criteria with correlations as follows:

State Boards Bacteriology	.31
Sociometric Research Promise	.21
Interest in Public Health	.20
Six National Boards Scores	.25 to .32
Intern, Over-all Medical Knowledge	.29
Intern, Diagnostic Competence	.31

By contrast, Strong VIB scores for Physician yielded but two significant correlations: -.31 with State Boards in Chemistry and Toxicology and -.21 with sociometric choice as Office Partner. The most likely explanation of the lack of validity of this score is that this group of subjects, both as the result of self-selection and the selective process of admission, was so relatively homogenous with respect to the interest pattern measured by the Physician key that there was little opportunity for covariance to occur.

Finally, the Anxiety score derived by scoring the Strong VIB responses with a specially developed key yielded a consistent array of positive correlations with nine criterion variables all heavily loaded with intellectual and academic accomplishment. Interestingly enough, this variable seems to be tapping something different than the Nervous Tension factor of Cattell 16 PF which was more likely to be positively correlated with sociometric choices.

Discussion

In view of the known uniqueness of many of the criterion measures employed, it is encouraging to discover that so many of the predictor variables showed significant and often meaningful correlations with so many criteria. Obviously, however, any practical program of student selection would require some consensus on the part of a faculty regarding the relative importance and hence the manner of weighting alternative criterion measures before making a decision regarding predictor variables. Fortunately, factor analyses of our criterion variables indicate that there are probably not more than five or six really meaningful dimensions involved. *If* satisfactory measures of this limited set of criteria could be developed, it is highly probable that even better predictive devices could be developed than the ones used in this study, e.g.—pre-med. grades might well be weighted by the median SAT score of freshmen admitted to each pre-med. college; the parts of the MCAT could be designed to predict more specific criteria, empirically derived keys for the Strong VIB might well provide more useful scores than the occupational keys now available, etc.

Obviously, however, no test or test battery and no statistical technique can answer the fundamental question of what kind of a physician the faculty of a given medical school wishes to produce. Given a multidimensional criterion, which appears to be the case, and relatively non-overlapping sets of predictor variables for each, the "yes-no" decision required in student selection must in the long run depend on the hierarchical ranking and weighting of the criterion dimensions by the faculty concerned.

The findings of the study here reported strongly suggest that wise decisions regarding the product desired cannot be arrived at by any amount of staff discussion of the problem in the abstract. Only with the aid of an ongoing program which monitors the characteristics of applicants selected and rejected and of the criteria used to assess success in both the school and in practice, and feeds back the interrelationships among these variables, will a staff be in the position of knowing to what degree its stated objectives are being attained. Fortunately, with the ready availability of modern computers, such an ongoing program of "quality control" is now entirely feasible for any medical school. Obviously, at least one appropriately trained professional person is needed to identify the essential variables, to collect and analyze the data in a systematic fashion, and to interpret the results back to the faculty members concerned. (11)

While the findings of this study of a single class in one medical school do not justify any recommendations regarding specific procedures to be used in the selection of medical students, they do point to a number of more general conclusions, each with implication for measurement in all institutions involved in professional training:

A. The criterion problem is both important and complex. Instead of a neat unidimensional criterion, it appears likely that there are several relatively unrelated criterion dimensions of success in professional education and practice. Since each of these dimensions is likely to be regarded as important by subgroups of the faculty and by segments of the society which the profession serves, it is essential that improved measures of these criterion dimensions be developed.

B. It appears likely that reasonably valid predictions of alternate criteria of professional performance can be made on the basis of data obtainable before admission to the professional school, but a different subset of predictor variables will obviously be required to predict uncorrelated criteria.

C. In view of the limited number of applicants for professional education (an applicant to medical school currently has better than one chance in two of being accepted by *some* medical school within a couple of years) and by virtue of the differential pattern of predictor variables correlated with alternate criteria of performance, it is simply not feasible for any school to attempt to select applicants who will rank high on all criterion dimensions. This suggests the possible desirability of an explicit decision on the part of the staff of each professional school with respect to the particular dimension(s) of professional performance which that school wishes to maximize both in its program of student selection and its program of instruction. Alternatively, larger professional schools may wish to consider the establishment of clearly differentiated programs of professional training with the consequent implications for using different variables in student selection and expecting very different kinds of professional performance in the graduates of the alternative programs of education.

REFERENCES

1. Flexner, A. "Medical Education in the United States and Canada." A report to the Carnegie Foundation for the Advancement of Teaching, Bulletin No. 4, New York, 1910.
2. Moss, F. A. "Aptitude Tests in Selecting Medical Students." *Personnel Journal*, 1931, 10, 79-94.
3. Kelly, E. L. "An Evaluation of the Interview as a Selective Technique." *Proceedings of the* 1953 *Invitational Conference on Testing Problems*, Princeton, New Jersey: Educational Testing Service, 1953.
4. "A Critique of the Interview" in *The Appraisal of Applicants to Medical Schools.* Evanston, Illinois: Association of American Medical Colleges, 1957.
5. Kelly, E. L. and Fiske, D. W. *The Prediction of Performance in Clinical Psychology*. Ann Arbor: University of Michigan Press, 1951.
6. Cattell, R. F. The Sixteen Factor Personality Questionnaire. Champaign, Illinois: Institute for Personality and Ability Testing, 1957.
7. Strong, E. K. The Vocational Interest Blank. Stanford, California: Stanford University Press.
8. Garman, G. D. and Uhr, L. "An Anxiety Scale for the Strong Vocational Interest Inventory: Development, Cross Validation and Subsequent Tests of Validity." *Journal of Applied Psychology*, 1958, 42, 241-246.
9. Tussig, L. "An Investigation of the Possibilities of Measuring Personality Traits with the Strong Vocational Interest Blank." *Educational and Psychological Measurement*, 1942, 2, 59-74.
10. McQuitty, L. L. Health Index 46 and 47. Mimeographed edition supplied by the author.
11. Kelly, E. L. "Multiple Criteria of Medical Education and Their Implications for Selection." *The Appraisal of Applicants to Medical Colleges*, Evanston, Illinois, 1957.

LUNCHEON ADDRESS

Growing

Jerome S. Bruner,
Center for
Cognitive Studies,
Harvard University

Will you forgive me if I use this as an occasion to clear my own mind on several issues that relate to the growth of intellect in human beings. I have been engaged these last several years in research on this opaque topic, reading, experimenting, puzzling over recalcitrant data, arguing with wiser men than I, even puzzling over the ancient problem of the parallelisms that may exist between the emergence of the species *Homo* and the growth of the child. The urge to clear my thoughts is not only generic, but rather specific in this case, for very shortly we shall be taking possession at the Center for Cognitive Studies of a quite handsomely equipped mobile laboratory that will make it possible for us to go out to where children are and test their functioning under standard conditions that, until now, have been hard to obtain. So if at times my conjectures may seem tortured or perhaps foolish, you will know that at least the motive is honorable and practical.

I should like to talk about several subjects that are particularly bedevilling, the first of which has to do with the nature of thought. I shall take it that by thinking we mean that an organism has freed itself from domination by the stimulus, that he is able to maintain an invariant response in the face of a changing stimulus input or able to vary response in the face of an invariant stimulus environment. In short, we can conceive of something remaining the same in some essential respect, though its appearance changes drastically, and also entertain different hypotheses about it though its appearance stays the same.

The means whereby an organism effects this freedom from stimulus control is through mediating processes, as they have come to be called in recent years. A mediating process consists at very least of some representation or model of the environment plus some rules for performing transformations on the representations such that the organism can represent not only past and present states, but also states of the world that might exist. Or, to use another set of words, thinking involves constructing a model of the world as we have experienced it and of having rules for spinning the model into positions from which we can read off predictions of things to come or things that might be. If all of this apparatus is to have any functional significance for the organism, then there must first of all be some correspondence between the model one constructs in one's mind (to use the old-fashioned term) and the world in which one must operate. And moreover, if the rules for spinning or transforming the model are to have any predictive or extrapolative value, they must also have some bearing upon the processes that go on in nature. What assures functional utility of this kind is, of course, feedback and correction that occur when we attempt to use thinking to deal with some domain of experience or potential experience.

There are various questions that immediately pose themselves, given this conception, and on closer inspection they turn out to be questions not only about the operation of thought but also about the nature of intellectual growth. Let me set these questions out, and then we can turn our attention to them seriously.

The first has to do with the nature of representation. How do we in fact represent the world? In this case I shall define the world simply as the recurrent regularities in man's experience and align myself with Ernst Mach (1914) in order to avoid any metaphysical fidgeting. The question of how to represent things turns out not only to be a problem for the psychologist interested in thought or memory, but also a problem for the computer simulator who faces the issue of how to organize storage and retrieval of information. For us the question becomes one about the development of representations. They change with growth. How?

Secondly, what accounts for the scope and connectedness of a particular representation? Some representations take in great generic chunks of the world, and permit ready recognition of the relations between things. Others are highly specific, event-bound and time-bound, almost assuring there will be very little transfer of knowledge and skill from one situation to another. This transferability increases enormously with growth. How does it come about?

Thirdly, how do we operate upon our models or representations of the world in order to predict or extrapolate or otherwise go beyond the information given? Are these operations like the laws of logic or of language or what? Surely they are not the same for all ages, for all conditions, that is plain—or else there would be no disagreements. If we have learned anything from Piaget (1950) at all it is certainly that the "logic" of the child of three is not that of the child of six. But you may well knit your brows over my use of the word logic, with or without quotation marks. For surely the operations of thought are not really "logic."

And finally, what is the role of the genetic code in the growth of man's capacity to construct and use models of the world? No matter how replete Aristotle's genetic code might have been, it did not contain information that made him able to deal with quadratic functions. Far less gifted mathematics concentrators in Harvard College today do it much more readily with less genetic code to go on. So perhaps it would be better to ask the question in the reverse direction. To what extent does the working out of man's genetic code, that part of it having to do with intellectual capacity, depend upon instruments, appliances, formulae, and other intellectual prosthetic devices? The issue in its barest form is rather startling upon reflection, for its resolution governs how we conceive of instruction, curricula, and the other means whereby we equip human beings to grow.

How do human beings construct models of their world and how do these change with growth? Second, how do these models become sufficiently general so that they fit a wide variety of situations we encounter? Third, how do we use models to go beyond the information given under our noses? And finally, what has all this to do with inheritance?

Now we can turn back. What is meant by *representation*? What does it mean to translate experience into a model of the world? Let me suggest that there are probably three ways in which human beings accomplish this feat. The first is through action. We know many things for which we have no imagery and no words and they are very hard to teach to anybody by the use of either words or diagrams and pictures. If you have tried to coach somebody at tennis or skiing or teach a child to ride a bike, you will have been struck at the wordlessness and the diagrammatic impotence of the teaching process. (I heard a sailing instructor a few years ago involved with two children in a shouting match about "getting the luff out of the main"; the children understood every single word, but the sentence made no contact with their muscles. It was a shocking performance, like much that goes on in school.) There is a second system of representation that depends upon visual or other sensory organization and upon the use of summarizing images. We may, as in an experiment by Mandler (1962), grope our way through a maze of toggle switches, and then at a certain point in overlearning, come to recognize a visualizable path or pattern. In Cambridge, we have come to talk about the first form of representation as *enactive*, the second as *ikonic*. Ikonic representation is principally governed by principles of perceptual organization and by the economical transformations in perceptual organization that Attneave (1954) has described – techniques for filling in, completing, extrapolating. Enactive representation is based, it seems, upon a learning of responses and forms of habituation.

Finally, there is representation in words or language. Its hallmark is that it is symbolic in nature with the design features of symbolic systems that are only now coming to be understood. Symbols (words) are arbitrary (as Hockett [1959] puts it, there is no relation between the symbol and the thing so that *whale* can stand for a very big creature and *microorganism* for a very small one), they are remote in reference, and almost always highly productive or generative in the sense that a language or any symbol system has rules for the formation and transformation of sentences that can turn reality over on its beam ends beyond what is possible through actions or images. A language, for example, permits us to introduce lawful syntactic transformations that make it easy and useful to approach declarative propositions about reality in a most striking way. We observe an event and encode it—the dog bit the man. From this utterance we can travel to a range of possible recodings—did the dog bite the man or did he not? If he did not, what would have happened, etc., etc.? Grammar also permits us an orderly way of stating hypothetical propositions that may have nothing to do with reality—"The unicorn is in the garden"; "I speak a triangle, a mystery"; "In the beginning was the word."

I should also mention one other property of a symbolic system—its compactability—a property that permits condensations of the order F=MA or S=1/2 gt^2 or "Gray is all theory/Green grows the golden tree of life," in each case the grammar being quite ordinary, though the semantic squeeze is enormous. My colleague George Miller (1956) has proposed a magic number 7±2 as the range of human attention or immediate memory. We are indeed limited in our span. Let me only suggest here that compacting or condensing is the means whereby we fill our seven slots with gold rather than dross.

Now what is abidingly interesting about the nature of intellectual development is that it seems to run the course of these three systems of representation. But one must say this more carefully. The young infant appears to operate by a process that is notably restricted to action, and there is very poor definition of the nature of objects "outside." This is not the occasion for reviewing research, but I would like to give a brief account of some work. To put it in a word, it is as if the objects of the world did not exist autonomously of the actions directed toward them by the child. Piaget (1962) has another nice demonstration of how the identity of the object depends upon action. If a very young child, 9 months old, has hold of a desirable object and the object is removed, the result will be screams. If the object is removed before it is held, the child will not mind. A little later, movement of the hand toward the object suffices to identify it and if it is removed while he is reaching, screams of protest result. Finally, the child will protest if the object is removed once it has been fixated visually, and so on What the child masters in that first year or 18 months is some way of giving the object identity short of actually having it muscularly in hand or of orienting to it muscularly. It is a limited world, the world of action.

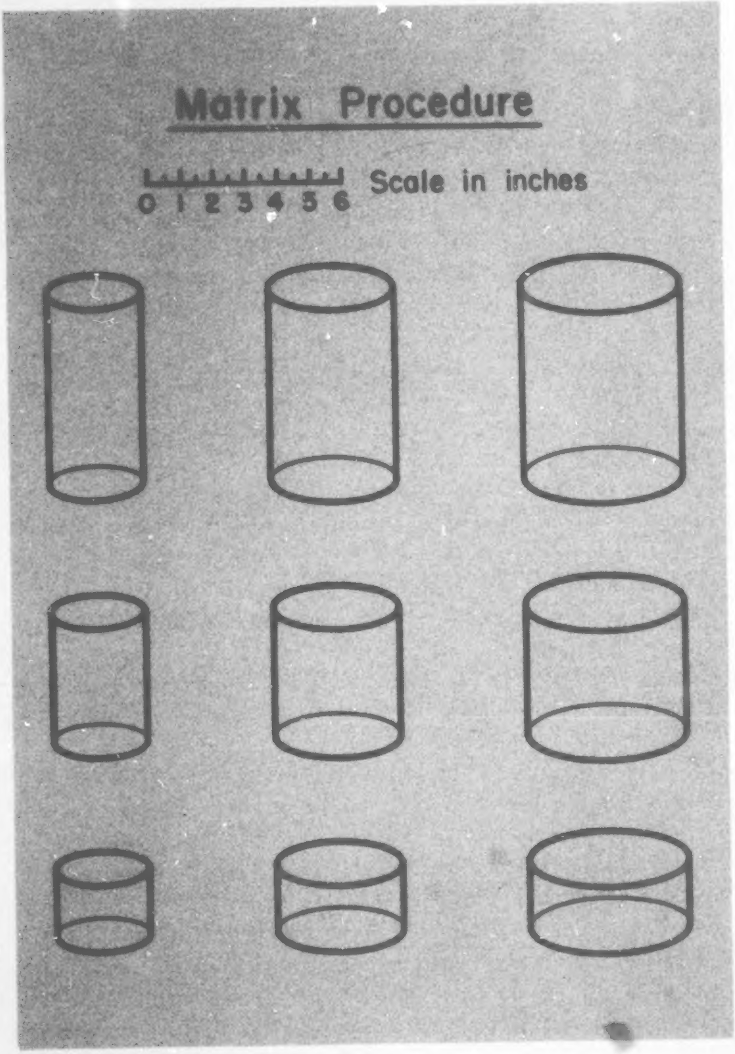
What appears next in development is a great achievement. Images develop an autonomous status, great summarizers of action. By age three the child has become a paragon of sensory distractibility. He is victim of the laws of vividness and his action pattern is a series of encounters with this bright thing which is then replaced by that chromatically splendid one, which in turn gives way to the next noisy one. And so it goes. Visual memory at this stage seems to be highly concrete and specific. What is intriguing about this period is that the child is a creature of the moment, as if the image of the moment is sufficient and it is controlled by a single feature of the situation. The child can reproduce things that were there before—in the form that was there before. He can reproduce a pattern of nine glasses laid out in rows and columns with diameter and height varying systematically (Figure 1). Indeed, he does it as well as a seven-year-old. But just change the position of one glass in the matrix that he has to reproduce and he is lost. He can copy from image but he cannot transform the image by transposition. The seven-year-old, on the other hand, can do it quite easily.

The difference seems to be a matter of being able to translate the visual experience into a form that can be operated upon, and here is where language is such a superb instrument of thought. For once the child is able to instruct himself in the task by saying to himself that in one direction the glasses get fatter and in the other they get taller, he can change the position of the matrix quite easily and without regard to orientation.

The child, of course, *has* language in nearly its full grandeur

FIGURE 1

Array of Glasses Used in Study of Matrix Ordering


Matrix Procedure
0 1 2 3 4 5 6
Scale in inches

by the time he is five—has it in the sense of using it in communication. But this is not the same as using it as an instrument of thought. I do not know quite how to say what it means to have language and use it as an instrument of thought as compared to having it and not using it in this way. But I rather suspect it has something to do with a process whereby the child, to use Sir Frederic Bartlett's (1950) old phrase, turns around on himself and reformulates what he does in a new form. Recall the subjects in the toggle-switch maze reformulating their way through the maze into a simultanizing image rather than representing it only by a successive series of gropings with minimum visual support. So too with language we seem to turn around on experience, reformulate, and condense it into language. Then we can use the transformative process that language makes possible.

Let me turn now to the issue of the scope of our models and their generic or transferable properties. How does the child learn to group experience into longer chunks so that it takes in longer periods of time and permits one to escape from immediacy? We find an interesting answer to this in our studies of the growth of inference. One of the principal features of growing up intellectually is being able to deal with indirect information. Let me illustrate. The young child has little success at the Twenty Questions game (at, say, age five) because he requires direct information, information that is self-sufficient. Why did a car bump into a tree? The five-year-old is full of direct and immediate tests of this or that hypothesis. A constraining, indirect strategy is beyond him: "Was it night?" "Yes." "Was anything wrong with the car?" etc.

Around seven, he comes to master the use of such strategies. It is interesting that at just about this time the child is also going through two parallel developments. On the one hand, he is learning to create rules of equivalence that join together a set of objects by what logicians speak of as a superordinate rule: that things may be considered alike because *all* of them exhibit a common characteristic. Before that, equivalence is not the true equivalence of the adult. Banana, peach, potato, milk are eventually all alike because they are all for eating, etc. But before that, banana and peach were alike because they were both yellow, peach and potato both have skins, peach and potato and milk I had for lunch yesterday. This latter equivalence rule is what Vygotsky (1962) years ago called complexive thinking, and we have reproduced his findings and have been able to write rules for such groupings. They are fantastically complicated rules in the sense that if you gave them to a computer, following them would demand very considerable memory and processing capacity. All such rules deal with local likeness in appearance—chains, keyrings, and so on. The passage to subordinate grouping provides a kind of freedom from the immediacy of local similarities. The other parallel development is the growth of the distinction in the child's thought between appearance and reality. Pour water into a standard glass. Then pour it from there into a slimmer, taller one. The child of five will say that the second glass had more water because it is taller. The child reckons by appearance. At seven, the picture changes. The child will say that it is the same amount of water to drink *really*, but it *looks* bigger.

Superordinate equivalence, appearance-reality distinction, and capacity to deal with indirect information—all within a year or so. We find, moreover, that the child can be aided to achieve this new simplicity by techniques that activate use of language before he encounters visually the real objects he is to deal with. We get him to talk about how things will be while the objects are hidden behind a screen and then expose him to them. The results are striking. The new system of representation by language seems to be able to compete under these conditions with the laws of image representation which contain no such distinctions of equivalence or of indirect information.

The growing scope of human "models" depends probably upon the opportunity for recoding experience into a language system that contains distinctions like those we have been discussing. How the language "gets into the head from the mouth," to quote a student of mine, is baffling in its details. But it may well depend upon some sort of law of intervening opportunity. Here is where the issue of assisted versus unassisted growth becomes central.

Language and the opportunity to use language in a fashion that is (in Dewey's lovely phrase) "a way of organizing thoughts about things," probably develops in interaction with an informal tutor—a parent or some adult member of the linguistic community who responds contingently to the child's responses by recoding or demanding a recoding. My colleague Roger Brown (in press) has shown the manner in which, in learning the syntactical structure of a language, the child first uses highly telegraphic utterances ("mummy coffee") which the parent then expands and idealizes to provide the child with a model ("Yes, mummy is having some coffee."). There are, very likely, games of this order that we quite unwittingly play with the child, and we know precious little about them.

But there are probably other things than language and symbolism that operate here. I have tried in vain to find something in the literature that is reliable on how a child scans his environment, whether he has uneconomical techniques for getting information. We have had to start experiments on our own. We have tried to find something about the child's immediate memory span or attention span. How many things can he hold in mind at once or, to use current jargon, what is his "channel capacity"? Again the literature is moot, so we shall put our mobile lab to work. But each of these things may be critically important. If the child's information search in the visual field is informationally inefficient (as we suspect it is), he will overload himself with too much material to cope with. If he cannot deal simultaneously with several alternatives, then again he cannot deal with equivalence problems which require that one carry over a criterion of grouping through several different items of apparent diversity. I wish this were two years from now so that I could vouchsafe a guess on this matter. My colleague George Miller would probably argue that there is nothing in this, that information capacity is probably not variable but is rather a matter of developing structures such that the Magic Number 7 ± 2 is filled with purer and purer gold.

What can be said about the "logic" that operates in thought during the sway of enactive, ikonic, and symbolic representation?

Here I am going to by-pass the issue altogether and say that far more research is needed. But I would like to hazard a guess that may serve for the moment as a working hypothesis. I think that at the earliest enactive phase, the principles of organization are the classic laws of frequency, recency, and proximity. What "goes together" as a model is that which has produced recent, frequent, and next-to responses. Probably Guthrie's (1952) psychology of learning or Pavlov's is the best description of early infancy. Inhibition at this stage depends upon stopping behavior by setting up a competing response. Learning is slow, gradual, and statistical. At the ikonic level, I would guess that the principles of figure formation and perceptual grouping determine the manner in which events are put together. It is the logic of appearances. The possibility of change depends upon perceptual reorganization, getting things to look different—which is swift and rather erratic in its effects. It would be foolish in the extreme to assert that the rules of language usage or empirical logic or any other such thing dominate the forming and reforming of models of experience in the symbolic phase of representation. For the fact of the matter is that at this stage there is enormous flexibility. My guess about the rules of thought when language takes over is simply to remain moot and observe. What I rather guess is that it is here that instruction becomes critically important. *How* the child uses symbolic representation in thought is partly a function of what behavior he has turned back upon to recode and upon the power and complexity of the rules that he has learned to use in this reflective process.

Man's history as a species suggests that there have not been any interesting and certainly no major morphological changes in man for some hundreds of thousands of years. As I have already suggested, he has progressed by linking himself with outside systems—evolution becomes alloplastic rather than automorphic, to use the technical language. Man's survival as a species, then, depends upon his flexibility in using means for amplifying his muscle power, his senses, and his ratiocinative capacities. As Peter Medawar (1963) has recently put it, evolution after the invention of a linguistic tradition becomes Lamarckian and reversible—but not as these terms were originally understood. Evolution achieves this new status by virtue of operating outside the genetic code. The further evolution of the species, then, depends upon the extent to which, in the development of each successive generation, there is mastery of the intellectual prosthetic devices of the culture. If a given generation succeeds well, then the next generation climbs on its shoulders. In an ironic vein, we can turn Haeckel's formulation on its head. Where human evolution is concerned, it is the case that phylogeny recapitulates ontogeny rather than vice versa.

With this in mind let me suggest one further point. Might it not be the case that the unlocking of the human genetic code (or that part having to do with intelligence) depends upon the invention of new amplifiers of human powers—new prosthetic devices, if you will? Because human evolution in the morphological sense seems to have determined the species as tool users (tools both hard and soft), we shall never know the full capacity of man until tool-using reaches the highest point it can reach—and here I mean those most powerful tools of all, intellectual ones. But might it not also be the case that the most interesting thing about intellectual tools is that their principal use is not print-out into technology alone, but that they make possible the creation of or the mastery of even more powerful tools? In this sense, evolution progresses by a system of prerequisites that is quite familiar to the teacher in us.

I end then with the paradoxical note that what we know about human growth suggests that education and the trained use of mind constitute our major agents for further evolution. In a group such as this, I can only add one point to make the conclusion directly relevant. One thing that has not been sufficiently part of the objective of testing is to discern what the far limit of man's capacities is at any given time—particularly during the fast developing years of childhood. Vygotsky (1962) commented years ago that perhaps it would be a good idea if we tested in children the so-called "zone of potential intelligence" —how much a child can make of the best hints we might give him, the best trots, the best tools, the maximum theoretical props and formulae. I am being deadly serious when I suggest that rather than testing under neutral conditions, we test under the most optimum conditions possible. To what extent, we should be asking in our tests, are the schools and the other agents of education using this child's capacities? Our present philosophy of testing too often asks only about aptitude and achievement. I would be delighted to see a year given over to teaching-and-testing-and-teaching-and-testing to see how far children can be brought along the Lamarckian way. Then and only then can testing serve us with benchmarks of not where a child *is* but where he is capable of *going*. And when we have fully exploited where *he* the individual is able to go, then we will be in a position to estimate where the species might go.

REFERENCES

Attneave, F. "Some Informational Aspects of Visual Perception." *Psychological Review*, 1954, 61, 183-193.

Bartlett, Sir F. C. *Remembering*. Cambridge: Cambridge University Press, 1950.

Brown, R. & Bellugi, Ursula (Eds.). "The Acquisition of Language." *Child Development Monograph*, in press.

Guthrie, E. R. *The Psychology of Learning*. (rev. ed.) New York: Harper, 1952.

Hockett, C. F. "Animal 'Languages' and Human Language." In J. N. Spuhler, *The Evolution of Man's Capacity for Culture*. Detroit: Wayne State University Press, 1959. Pp. 32-39.

Mach, E. *The Analysis of Sensations*. (Trans. from First German Ed. by C. M. Williams, revised and supplemented from Fifth German Ed. by S. Waterlow) Chicago & London: Open Court, 1914.

Mandler, G. "From Association to Structure." *Psychological Review*, 1962, 69, 415-427.

Medawar, P. "Onwards from Spencer: Evolution and Evolutionism." *Encounter*, 1963, 21(3), 35-43 (September).

Miller, G. "The Magical Number 7, Plus or Minus Two: Some Limits on our Capacity for Processing Information." *Psychological Review*, 1956, 63, 81-97.

Piaget, J. *Le développement des quantites physiques chez l'enfant*. Neuchatel, Switzerland: Delachaux et Niestlé, 1962.

Piaget, J. & Inhelder, Bärbel. *The Psychology of Intelligence*. (trans. by M. Piercy & D. Berlyne) London: Routledge & Kegan Paul, 1950.

Vygotsky, L. *Thought and Language*. New York: MIT Press & John Wiley, 1962.

Session III

Theme: Implications and Consequences of Measurement

Ability and Performance

Warren G. Findley,
College of Education,
The University of
Georgia

Over the years, the types of individuals represented at this conference have devoted untold hours to exploring the nature of ability. From time to time a new line of investigation or a new research technique has seemed to promise a breakthrough to some sort of fundamental truth or understanding regarding the organization and development of ability. (Kornhauser's 1944 questionnaire to 79 specialists in mental tests found 55 more hopeful of research on separate intellectual factors than on measurement of "general" intelligence, with only five committed to the opposite view.) At other times it has seemed that evidence from different sources was conflicting, if not contradictory or irreconcilable. The notion that "you get out of such a study just about what you put into it" has often seemed both true and discouraging for those who had hoped definitive findings could clear the scene of prevailing confusions.

Today, however, there seems to be emerging a possibility of reconciliation based on what might be called multiple or pluralistic truth. Physicists have learned to live comfortably for a generation or more with the fact that some of the behavior of light is well described by wave theory while other phenomena of this area are better explained by a view of light as the behavior of corpuscles moving under laws that fit the behavior of billiard balls equally well. The reconciliation in our field of mental ability is following a hierarchical view of the nature of ability.

Under this view, attributable chiefly to Vernon, but clarified greatly by Humphreys' delineation of the relation between orders of factors and levels of the hierarchy, an interpretation like Spearman's of a single primary intellective energy, *g*, surrounded by satellite uncorrelated specific factors, may represent the truth at the most general level of coping intellectually with the demands of the environment. At a second level, intellectual ability is represented by the two constellations we have come to call variously verbal and nonverbal, language and nonlanguage, or verbal and performance.

At a third level, these constellations may be subdivided further. What has been called verbal at the second level subdivides into verbal and quantitative reasoning factors. (Perhaps the verbal category at the second level is better called academic or scholastic ability, to conserve a term, especially since the most direct derivation of "verbal" is from "word" rather than from a more inclusive source.) At the same time, the performance or nonverbal constellation breaks down into spatial, mechanical, perceptual factors.

At still other levels we find the primary mental abilities of Thurstone and his followers. At what we should probably call the penultimate level, we have Guilford's structure of intellect with its instructive taxonomic uses. I say "penultimate" because one must conceive the possibility of still further refinement. To turn to science again for a helpful parallel, successively finer subdivisions of matter below the atoms that were once defined as the ultimate units of matter have given us power of which we could only have dreamed. If the structure of intellect may be thought of as the periodic table of abilities, vive les isotopes!

The catholicity of this viewpoint is even greater than has already been suggested because it leaves room for different causal interpretations of the different orders of factors. If one is impressed as J. McV. Hunt is by the analytical work of Piaget regarding the development of general strategies of thinking in children, by the speculations of Hebb regarding the development of intellect by stimulation and elaboration of the central neural processes that intervene between the sensory and motor, and by the logic of Ferguson, the factor-theorizing of Humphreys, and the factoring of growth data by Hofstaetter, he may prefer the notion of general intelligence advanced by Thomson and E. L. Thorndike over 50 years ago as a composite of many autonomous, but interacting and correlated skills. The behavior at the general level of such an intelligence and Spearman's *g* will not be statistically differentiated, so both may be accommodated until more fundamental experimentation can resolve the issue. One with this viewpoint may go as far as Piaget to disregard entirely intra-individual trait variability and concentrate on just the general level in the hierarchy. He may consider factors of lower order just that, relatively unimportant, as Hunt seems to. He may, like Humphreys, reject the notion of factors as primary mental abilities and simply think of them in descending order of significance. Or one may go all the way with J. P. Guilford in ascribing possible dynamic significance to the factors of the structure of intellect.

Two personal comments: Your speaker would concur in Chronbach's view of the importance of criterion-oriented tests like the Differential Aptitude Tests, which tend to fall at the third level in the hierarchy as I have described it. And on further review, I believe I may have inserted that level into a model that otherwise moved from the level of language-nonlanguage directly to the primary mental abilities. For I have always felt that, like the D.A.T., the Scholastic Aptitude Test of the College Entrance Examination Board is to be found at this level. Verbal reasoning ability and quantitative reasoning ability are basic academic skills, relating reasoning power to two functionally distinct media in the academic curriculum. They were identified by factor analysis quite early, by Brigham and T. L. Kelley, and they remain functional unities. For school purposes they are predictive and more meaningful than a scheme of three primary mental abilities: a large reasoning factor and two disembodied factors of trivial skills of numerical manipulation and verbal association. And because they merge reasoning and relevant skills in power tests, they are predictive of achievement in definable sub-segments of the academic curriculum.

The second comment may be considered by some to be merely semantic, but it seems significant to your speaker. It has to do with the refinement of classification in the structure of intellect and is reflected in the composition of the kit of reference tests offered to experimenters in the factor analysis of new tests or other measures in the cognitive domain. Both the primary mental abilities and the structure of intellect relegate skill in mathematics to a low place at table (or in the table?), something more characteristic of bookkeepers than of mathematicians. Yet in the structure of intellect and the kit of vector tests, "general" reasoning ability is represented by four tests, *all* of which are mathematical reasoning tests! It is quite possible to reach this position by adherence to factorial purity, but the relation between tests of mathematical reasoning ability at the third level of the hierarchy to mathematical measures of "general" reasoning ability at penultimate levels is worth pondering. One is tempted to ask which placement involves the more parsimonious description.

A corollary feature of the hierarchical model is that it allows lower order factors to be used either for their own sakes as significant entities at an appropriate level of understanding or using mental ability factors, or as guides to proper balance in the measurement of mental ability factors of a higher order. For example, the verbal and quantitative reasoning factors that emerged first from rudimentary methods in the years B. T. (before Thurstone) were used by McNemar and associates in redressing the balance in the Stanford-Binet. Criticisms of the 1916 version of that battery had included the observation that at its lower age levels verbal items and exercises predominated, while at the upper levels quite as great a predominance of quantitative elements was to be found. Analysis by factor methods showed the extent of this disparity statistically and was used to guide the choice of elements at all levels in the 1937 and subsequent revisions. The Stanford-Binet still yields a single score for general mental ability, but with due regard for balance within the sub-areas measured at the third level. If there is any question raised now it is probably for failure to take adequate account of the two major areas at the second level, language versus nonlanguage, as the Wechsler measures do explicitly.

To turn for the moment to current methods of appraising achievement in schools and colleges, let us pay tribute to the fundamental contribution of Ralph Tyler and his associates for their work in the 1930's in breaking the mold of testing for static, encyclopedic knowledge that characterized the first wave of achievement tests and batteries that had been developed in the preceding decade. The companion contribution of item-styles that measured higher mental processes through multiple-choice forms must be reckoned of equal significance. By 1942, concomitant thinking of Lindquist had led to construction of the Iowa Tests of Educational Development, which were ready to serve the important purpose of the United States Armed Forces Institute in measuring readiness of returning scholar-soldiers for college work or, at least, high school credit. (It should be noted here that earlier traces of this approach were to be found in the Iowa Every-Pupil Tests of Basic Skills, for grades 3-9, and in the Cooperative Achievement Tests being developed under Flanagan.)

The USAFI Tests of General Educational Development were work-limit tests. With their omission of time limits, they set a realistic miniature of the school study situation and thereby provided a yardstick against which in subsequent test building many agencies could accept the concept of power tests with generous time limits permitting many students to finish early. The time limit now became a means of assuring most examinees an opportunity to give as much time as necessary to complete the power-graded materials, rather than a uniform time in which to accomplish as much material—often of only moderate difficulty—as possible in a time in which only the most competent and facile could hope to finish.

The trend we hailed of shifting the emphasis in achievement testing from memoriter knowledge to ability to apply such knowledge has an opposite source of concern. In preparing tests to ascertain whether examinees can apply knowledge, some have gone so far as to remove in large part any requirement that the examinees draw upon a background of well-structured, important knowledge in answering the questions posed. A generally bright person with ability to interpret verbal, quantitative and graphic material may obtain a creditable score on such tests despite having failed to develop the systematic knowledge of the field we increasingly feel we should demand.

It is relevant to recall the experience of Kelley and Krey in the American Historical Association's 1934 report on their study of the teaching of the social studies. It had been intended to build a test depending entirely on ability to apply knowledge, but by mistake a factual item on the Sepoy Mutiny had been left in the test. When the many items of this test were correlated with total score, the item showing the highest correlation was the one inadvertently included. The explanation given was that the Sepoy Mutiny was such an important incident in the history of British colonial administration that better students, however defined, would be bound to remember it for that reason. Our ideal is, of course, a test in which both background and application are required in each item.

The USAFI Tests of General Educational Development afford a natural bridge to our third topic of relating ability to performance. It is the thesis of this paper that the hierarchical view of mental ability stated first lends itself to an eclectic approach to the use of measures in predicting achievement in any given situation. At no point was ability defined as inherited or the product of inevitable processes of maturation. For that reason any measure predictive of likely success in any particularly defined intellectual arena is an appropriate measure of aptitude for that success. (Substitution of PLR, meaning Probable Learning Rate, for IQ in a number of school systems has semantic merit.) In the case of the USAFI GED Tests, their use with returning soldiers to predict fitness for college study was enhanced by the extent to which they suppressed the requirement of knowledge ordinarily available in systematic form from recent advanced study. In a situation in which systematic knowledge from recent study is available, that may well be drawn upon in testing or through the evidence of school grades to supplement testing that does not require it.

The use of measures of ability to predict performance has been subjected to illuminating systematic treatment by one of this morning's speakers, Dr. Thorndike. I have enjoyed the quote attributed to him in news releases regarding his recent monograph on "The Concepts of Over- and Underachievement" that the term "underachiever" should be reserved for those who conceived the terminology in the first place. After pondering the logical fallacy implied in the terms, namely that if an underachiever is one who has done less than he can do, an overachiever must be one who has done more than he can do, I can only suggest the following substitutes: (1) We are all underachievers, only some are more so; and (2) an overachiever is simply an under-underachiever.

The constructive view emerging from all of this would appear to be that particular learning situations place demands on particular combinations of intellectual abilities, that these abilities are of different orders of generality in the hierarchical structure, and that all measures of abilities are measures of achievement of the appropriate order of generality. As long ago as 1937, Bingham stated the case for achievement as the best predictor of further achievement. Wesman has presented the statement in brief persuasive form. Our problem then would appear to be to use that combination of measured abilities most descriptive of aptitude, i.e. – most predictive in particular situations.

It is therefore no departure from sound conceptualization to propose appraising reading comprehension relative to listening comprehension, as has been done for years in the Durrell-Sullivan Reading Capacity and Achievement Tests. Other predictive helps become appropriate in special situations. Generally, measures of verbal mental ability, at the second or third level, will be helpful. Generally, measures of performance mental ability will not be so helpful. On the other hand, where special language handicaps like bilingualism are involved, something nonverbal will have advantages.

In reviewing the manual for the current edition of the Metropolitan Achievement Tests recently, it was interesting to note that the measure proposed for evaluating learning potential is a "composite prognostic score" composed chiefly or entirely of the previous year's achievement measures. Justification is given in terms of the greater stability of this type of composite over a year (.90) than of a well-regarded group test of mental ability over the same period (.80). Other studies of the same test by staff members suggest the desirability of different predictive equations for different subjects, depending on the size of the correlation, and even on different combinations of subjects, chiefly dependent on the distinction between verbal and quantitative reasoning abilities of the sort distinguished at the third level in the hierarchical model proposed in this paper.

A final note, that does not fit well into the general framework of this paper, deserves mention. For some time, the demand that measures of ability to write effective prose compositions be included in standardized test batteries has been met with the response from test specialists that such writing cannot be measured reliably in the time ordinarily allotted to standardized testing in schools or in external testing programs. Some recent research indicates that global rating of pieces of writing on specified topics, with the reliability enhanced as far as feasible by multiple rating, will permit scores of adequate reliability and validity to be reported. We may well still counsel that cumulative evidence of writing ability be obtained by systematic evaluation of weekly compositions, but it seems to be becoming increasingly possible to appraise such outcomes within test batteries and thereby give comparable emphasis to this skill outcome along with others ordinarily appraised by objective tests.

REFERENCES

Bingham, Walter V. *Aptitudes and Aptitude Testing.* New York: Harpers, 1937.

Brigham, Carl C. *A Study of Error.* New York: College Entrance Examination Board, 1932.

Cronbach, Lee J. *Essentials of Psychological Testing* (second edition) New York: Harpers, 1960.

Durost, Walter N. *Manual for Interpreting Metropolitan Achievement Tests.* New York: Harcourt, Brace and World, 1962.

Ferguson, G. A. "Learning and Human Ability: A Theoretical Approach" in DuBois, P. H., et al (eds.) *Factor Analysis and Related Techniques in the Study of Learning.* Technical Report No. 7, Office of Naval Research Contract No. Nonr 816 (02), 1959.

Guilford, J. P. "The Structure of Intellect" *Psychological Bulletin*, 53, 267-293. 1956.

Guilford, J. P. "A Revised Structure of Intellect" *Reports of the Psychological Laboratory*, No. 19. Los Angeles: University of Southern California, 1957.

Hebb, D. O. *The Organization of Behavior.* New York: Wiley, 1949.

Hofstaetter, P. R. "The Changing Composition of 'Intelligence': A Study in T-technique" *Journal of Genetic Psychology*, 85, 159-164, 1954.

Humphreys, Lloyd G. "The Organization of Human Abilities." *American Psychologist* 17, 475-483, July 1962.

Hunt, J. McV. *Intelligence and Experience.* New York: Ronald Press, 1961.

Jenkins, James J. and Paterson, Donald G. *Studies in Individual Differences.* New York: Appleton Century Crofts, 1961.

Kelley, T. L. *Crossroads in the Minds of Man.* Stanford: Stanford University Press, 1928.

Kelley, T. L. and Drey, A. C. *Tests and Measurements in the Social Sciences.* New York: Scribners, 1934.

Kornhauser, Arthur. "Replies of Psychologists to a Short Questionnaire on Mental Test Developments, Personality Inventories, and the Rorschach Test." *Educational and Psychological Measurement*, 5, 3-15, Spring 1945.

McNemar, Quinn. *The Revision of the Stanford-Binet Scale.* Boston: Houghton Mifflin, 1942.

Piaget, J. *The Psychology of Intelligence.* London: Routledge and Kegan Paul, 1947.

Smith, E. R. and Tyler, R. W. *Appraising and Recording Student Progress.* New York: Harpers, 1942.

Spearman, Charles E. *The Abilities of Man.* New York: Macmillan, 1927.

Thorndike, Robert L. *The Concepts of Over- and Underachievement.* New York: T. C. Bureau of Publications, 1963.

Thurstone, L. L. *Primary Mental Abilities.* Chicago: University of Chicago Press, 1938.

Tyler, Ralph W. *Constructing Achievement Tests.* Columbus, Ohio: Ohio State University Press, 1934.

Vernon, P. E. *The Structure of Human Abilities.* New York: Wiley, 1950.

Wesman, Alexander G. "What Is an Aptitude?" *Test Service Bulletin No.* 36. New York: Psychological Corporation, 1948.

Wesman, Alexander G. "Aptitude, Intelligence and Achievement" *Test Service Bulletin No.* 51. New York: Psychological Corporation, 1956.

Personality Measurement and College Performance*

SAMUEL MESSICK,
Educational Testing Service

In this paper I will discuss personality measurement primarily in terms of its potential contributions to the prediction of college performance. In this context, two major questions arise: (1) Are personality tests any good as measures of the purported personality characteristics? (2) What should these tests be used for? The first question is a scientific one and may be answered by an evaluation of available personality instruments against scientific standards of psychometric adequacy. The second question is at least in part an ethical one and may be answered by a justification of proposed uses for a test in terms of ethical standards and social or educational values. I will first discuss the scientific standards for appraising personality measures and will then consider how well these standards are typically met by instruments developed by each of three major approaches to personality measurement. The final section of the paper will discuss some of the ethical problems raised when personality measures are used for practical decisions.

---

*A preliminary version of some portions of this paper was prepared for the Committee of Examiners in Aptitude Testing of the College Entrance Examination Board. The author wishes to thank Dr. Salvatore Maddi for his many suggestions about the nature of the problems and the organization of the material. Grateful acknowledgment is also due Sydell Carlton, Norman Frederiksen, John French, Nathan Kogan, and Lawrence Stricker for their helpful comments on the manuscript.

Samuel Messick

Psychometric Standards for Personality Measurement

The major measurement requirements in personality, as in psychology generally, involve (1) the demonstration, through substantial consistency of response to a set of items, that *something* is being measured; and (2) the accumulation of evidence about the nature and meaning of this "something," in terms of the network of the measure's relations with theoretically relevant variables and its lack of relation with theoretically unrelated variables (Cronbach & Meehl, 1955; Loevinger, 1957; Bechtoldt, 1959; Campbell & Fiske, 1959; Campbell, 1960; Ebel, 1961). In psychometric terms, these two critical properties for the evaluation of a purported personality measure are the measure's *reliability* and its *construct validity.*

An investigation of the measure's relations with other well-known variables may also provide a basis for determining whether the thing measured represents a relatively separate dimension with important specific properties or whether its major variance is predictable from a combination of other, possibly more basic, characteristics. Such information bears upon the status of the construct as a separate variable and upon the structure of its relations with other variables.

Whether the measure reflects a separate trait or a combination of characteristics or, indeed, whether the proposed construct is a valid integration of observed response consistencies or merely a gratuitous label, there is still another important property of the measure that can be independently evaluated—namely, its usefulness in predicting concurrent and future non-test behaviors as a possible basis for decision making and social action. For such purposes, which primarily include classification and selection situations, it is necessary that the measure display *predictive validity* in the form of substantial correlations with the criterion measures chosen to reflect relevant performances in the non-test domain. Although some psychologists would argue that such predictive validity is all that's necessary to warrant the use of a measure in making practical decisions, it will be maintained here that predictive validity is not sufficient and that it may be unwise to ignore construct validity even in practical prediction problems (Gulliksen, 1950; Frederiksen, 1948; Frederiksen, 1954). This point will be discussed more fully later.

Just as a test has as many empirical validities as there are criterion measures to which it has been related, so too may a test display different proportions of reliable variance or reflect different construct interpretations, primarily because the motivations and defenses of the subjects are implicated in different ways under different testing conditions. Thus, instead of talking about the reliability and construct validity (or even the empirical validity) of the *test* per se, it might be better to talk about the reliability and construct validity of the *responses* to the test, as summarized in a particular score, thereby emphasizing that these test properties are relative to the processes used by the subjects in responding (Lennon, 1956). These processes, in turn, may differ under different circumstances, particularly those affecting the conceptions and intentions of the subjects. Thus, the same test, for example, might measure one set of things if administered in the context of diagnostic guidance in a clinical setting, a radically different set of things if administered in the context of anonymous inquiry in a reasearch laboratory, and yet another set if administered as a personal evaluation for industrial or academic selection. Furthermore, these different testing settings impose different ethical constraints upon the manner and conditions of eliciting personal, and what the subject may consider private, information ("Standards of Ethical Behavior, 1958"; Cronbach, 1960).

This point that personality tests, and even personality testers, may operate differently under different circumstances was one of the main reasons I initially chose to limit the present discussion to a particular context—namely, personality measurement in relation to college performace. Various contexts differ somewhat in the types of problems posed for personality measurement, but the timely context of assessment for college contains nearly all the problems at once. Of major concern in considering this context, however, is the inherently evaluative atmosphere of the testing settings. This means that we must take into account not only the ubiquitous response distortions due to defense mechanisms of self-deception and personal biases in self-regard (cf. Frenkel-Brunswik, 1939), but also the distortions in performance and self-report that are at least partially deliberate attempts at faking and impression management (cf. Goffman, 1959).

The extent to which attempts are made to handle the problems of both deliberate misrepresentation and unintentional distortion becomes an important criterion for evaluating personality instruments, particularly for use in evaluative settings. Many personality measures have been developed in research contexts where deliberate misrepresentation may have been minimal; little is known of their psychometric properties under conditions of real or presumed personal evaluation. Some personality tests include specific devices for detecting faking, such as validity or malingering keys, which would enable students with excessive "lie" responses to be spotted and would also permit the use of the control scores as suppressor variables in correcting other scales (Meehl & Hathaway, 1946). Other personality instruments rely on test formats that attempt to make faking difficult, such as the use of forced-choice techniques on questionnaires or of objective performance measures where the direction of faking is not obvious. Still other procedures use indirect items and disguised façades to circumvent the subject's defensive posture (Campbell, 1950; Campbell, 1957; Loevinger, 1955).

Psychometric Problems in Some Typical Approaches to Personality Measurement

We have considered several psychometric criteria for evaluating personality measures: reliability, empirical validity in predicting criteria or non-test behaviors, the structure of relations with other known variables, the adequacy of controls for faking and distortion, and—more basic because it subsumes aspects of the preceding properties—construct validity. We will now inquire how well these standards are typically met by instruments developed by three major approaches to personality measurement—

self-report questionnaires, behavior ratings, and objective performance tests.

SELF-REPORT INVENTORIES

Before various types of self-report questionnaires are discussed, the general problem of stylistic consistencies or response sets on such instruments should be broached (Cronbach, 1950; Jackson & Messick, 1958). A major portion of the response variance on many personality inventories, particularly those with "True-False" or "Agree-Disagree" item formats, has been shown to reflect consistent stylistic tendencies that have a cumulative effect on presumed content scores (e. g., Edwards, 1957; Edwards, Diers, & Walker, 1962; Jackson & Messick, 1961, 1962a). The major response styles emphasized thus far are the tendency to agree or acquiesce (Couch & Keniston, 1960; Messick & Jackson, 1961), the tendency to respond desirably (Edwards, 1957; Messick, 1960), the tendency to respond deviantly (Berg, 1955; Sechrest & Jackson, 1963), and, to a lesser extent, the tendency to respond extremely in self-rating (Peabody, 1962). These response styles have been conceptualized and studied as personality variables in their own right (Jackson & Messick, 1958), but their massive influence on some personality inventories can seriously interfere with the measurement of other content traits (Jackson & Messick, 1962b). The problem becomes one of measuring response styles as potentially useful personality variables and at the same time controlling their influence on content scores (Messick, 1962; Wiggins, 1962). The extent to which controls for response styles have been effective in reducing overwhelming stylistic variance becomes an important criterion in evaluating the measurement characteristics of self-report instruments.

We will consider three kinds of self-report or questionnaire measures of personality: (1) a type that I will call a *factorial inventory*, in which factor analysis or some other criterion of internal consistency is used to select items reflecting homogeneous dimensions (Cattell, 1957; Comrey, 1962); (2) *empirically derived inventories*, in which significant differentiation among criterion groups is the basis of item selection; and (3) *rational inventories*, in which items are chosen on logical grounds to reflect theoretical properties of specified dimensions.

- *Factorial inventory* scales are developed through the use of factor analysis or other methods of homogeneous keying (Wherry & Winer, 1953; Loevinger, Gleser, & Dubois, 1953; Henrysson, 1962) to isolate dimensions of consistency in response to self-descriptive items. The pool of items collected for analysis usually consists of a conglomeration of characteristics possibly relevant to some domain and sometimes includes items specifically written to represent the variables under study.

The most widely known of the current factored inventories are the Cattell 16 Personality Factor Questionnaire and the Guilford-Zimmerman Temperament Survey. Becker's (1961) recent empirical comparison of the Cattell questionnaire with an earlier form of the Guilford scales has revealed an equivalence between four factors from the two inventories and substantial similarity for two other factors. Although considerable factor analytic evidence at the item level generally supports the nature of the scales (Cattell, 1957; Guilford & Zimmerman, 1956), when two subscale scores were used to represent each factor supposedly measured by these inventories, Becker (1961) found only eight distinguishable factors within the 16 P. F. and only five within 13 Guilford scales.

These factorial inventories were developed primarily in research settings, so that attention must be given to possible defensive distortions induced by their use in evaluative situations. Although procedures for detecting faking have been suggested, their systematic use has not been emphasized, nor has their effectiveness been clearly demonstrated. Further, empirical controls for response styles have usually not been included, although their operation has recently been noted on some of the factor scales (Bendig, 1959; Becker, 1961).

- In the construction of *empirically derived inventory* scales, items are selected that significantly discriminate among criterion groups. The most widely known examples are scales from the

Minnesota Multiphasic Personality Inventory (MMPI) and from the California Psychological Inventory (CPI). The justification of these scales is in terms of their empirical validity and their usefulness in classifying subjects as similar or dissimilar to criterion groups. Scale homogeneity, reliability, and construct validity are seldom emphasized. The difficulty arises when these scales are used not to predict criterion categories but rather to make inferences about the personality of the respondent. This latter use has become the typical one (cf. Welsh & Dahlstrom, 1956), but such application cannot be justified by empirical validity alone—homogeneity and construct validity become crucial under such circumstances (Cronbach, 1958; Jackson & Messick, 1958, 1962b).

Because of their widespread use in clinical settings, considerable attention has been given to the problem of faking, particularly on the MMPI. Several scales are available for detecting lying and malingering (L, Mp, Sd, etc.), along with a validity scale (F) for uncovering excessive deviant responses (Dahlstrom & Welsh, 1960). A measure of "defensiveness" (K) is also used both as a means of detecting this tendency and as a suppressor variable for controlling test-taking attitudes (Meehl & Hathaway, 1946). Several studies of the effectiveness of these scales have indicated a somewhat variable, and usually only moderate, level of success (cf. Welsh & Dahlstrom, 1956; Wiggins, 1959).

A major problem on the MMPI and CPI is the predominant role of the response styles of acquiescence and desirability, which in the former instrument define the first two major factors and together account for roughly half the total variance (Jackson & Messick, 1961, 1962a; Jackson, 1960; Edwards, Diers & Walker, 1962). Presumably, these response styles are correlated with the criterion distinction utilized in the empirical scale construction (cf. Wahler, 1961), but their massive influence on these inventories drastically interferes with the attempted measurement of other content traits and limits their possible discriminant validity (Jackson & Messick, 1962b).

- *Rational inventories* comprise items that have been written on theoretical or logical grounds to reflect specified traits. That such scales measure something is demonstrated subsequently by high internal consistency coefficients; that they measure distinguishable characteristics is shown by relatively low scale intercorrelations. Factor analysis is also sometimes used subsequently to investigate scale interrelations (Stern, 1962). On some of these inventories, such as Stern's Activities Index, little attention has been given initially to the role of response styles, while on others, such as Edwards Personal Preference Schedule (EPPS), the major attraction has been the attempt to limit stylistic variance.

The EPPS employs a forced-choice item format: statements are presented to the subject in pairs, the members of each pair having been previously selected to be as equal as possible in average judged desirability. The respondent is required to select from each pair the statement that better describes his personality. Such forced-choice items do not offer an opportunity for the response style of acquiescence to operate. Further, since the paired statements are also approximately matched in desirability, a consistent tendency to respond desirably should in principle have relatively little effect upon item choices (Edwards, 1957; Corah et al., 1958; Edwards, Wright, & Lunneborg, 1959). Even though desirability variance is not eliminated thereby, primarily because of the existence of consistent personal viewpoints about desirability that cannot be simultaneously equated (Rosen, 1956; Borislow, 1958; Heilbrun & Goodstein, 1959; Messick, 1960; LaPointe & Auclair, 1961), the forced-choice approach offers considerable promise for reducing the overwhelming influence of response styles on questionnaires (Norman, 1963b). Unfortunately, the EPPS can still not be recommended for other than research purposes because insufficient evidence exists concerning its empirical and construct validity (Stricker, 1963).

The different approaches to scale construction that distinguish factorial, empirically derived, and rational inventories might well be combined into a single measurement enterprise, wherein scale homogeneity, construct validity, and the theoretical basis of item content, as well as empirical differentiation, would be successively refined in an iterative cycle (Loevinger, 1957; Norman, 1963b). In this way the differences among the approaches, depending as they would upon the particular point in the cycle that one chose to start with, would become trivial, and scales would be systematically developed in terms of joint criteria of homogeneity, theoretical relevance, construct validity, and empirical utility.

BEHAVIOR RATINGS

Behavior ratings represent a second major approach to personality measurement. Direct ratings of behavior, both of job performance and of personality characteristics, have been frequently employed in educational and industrial evaluation (Whisler & Harper, 1962). Personality ratings, however, have seldom been formally or systematically used in the typical selection situation for many reasons, one of them being the difficulty of obtaining reliable or comparable ratings for candidates coming from different sources. However, if teacher- and peer-ratings of personality made in college, for example, were to prove valid in predicting behavioral criteria of college success (cf. Tupes, 1957) and if these ratings could, in turn, be predicted by other measures (such as self-report inventories), then the predicted ratings might be useful in pre-college decisions. Behavior ratings that correlate with college success could thus serve as intermediate criteria for validating self-report measures of the same dimensions.

Cattell (1957) has isolated approximately 15 dimensions from behavior ratings, reflecting such qualities as ego strength, excitability, dominance, and surgency. Tupes and Christal (1961), on the other hand, in analyzing the same rating scales and in a few cases the same data, provided evidence for only five strong and recurrent factors, which were labeled extroversion, agreeableness, conscientiousness, emotional stability, and culture (see also Norman, 1963a). Cattell (1957) has also claimed a congruence between most of his behavior rating factors and their questionnaire counterparts, which suggests that questionnaire scales can indeed predict rating dimensions. Cattell's claim of a one-to-one matching of behavior rating and questionnaire factors has been challenged by Becker (1960), however, who concluded that available evidence did not support the alleged relation.

Norman (1963b), on the other hand, has clearly demonstrated that questionnaire scales can be developed that will correlate substantially with behavior rating factors. In his particular study, he attempted to predict the five rating factors obtained by Tupes and Christal (1958, 1961) from peer nominations. Since these ratings had previously exhibited substantial validity in predicting officer effectiveness criteria at the USAF officer candidate school (Tupes, 1957), the subsequent prediction of these ratings by questionnaire scales has direct implications for selection.

Incidentally, Norman's (1963b) scale construction procedure involved an extremely promising technique for handling faking in evaluative settings. Items in a forced-choice format equated for "admission-to-OCS desirability" were administered under normal and faking instructions. In the construction of the scales, the items were balanced between those showing a mean shift under faking instructions in the direction of the keyed response and those showing a mean shift away from the keyed response. Mean scores for the resulting scales were thus equated under normal and faking conditions, and, in addition, powerful detection scales were developed to isolate extreme dissemblers.

OBJECTIVE PERFORMANCES TESTS

The third major approach to personality measurement considered here is the objective performance test. According to Campbell (1957), an objective measure of personality, like an objective measure of ability or achievement, is a test in which the examinee believes that he should respond accurately because correct answers exist as a basis for evaluating his performance. Cattell (1957), on the other hand, considers a test objective if the subject is unaware of the manner in which his behavior affects the scoring and interpretation, a property that Campbell (1957) prefers to use in the definition of indirect measurement.

Cattell's (1957) analyses of objective performance measures of personality have uncovered approximately 18 dimensions, with such labels as harric assertiveness, inhibition, anxiety, and critical practicality. Thurstone (1944) and Guilford (e. g., 1959) have also developed measures of several perceptual and cognitive dimensions that represent objective tests of personality. Measures of speed and flexibility of closure (Thurstone, 1944), for example, and of ideational fluency appear more congenial in a personality framework than in the traditional ability formulation (cf. Cattell, 1957; Guilford, 1959; Witkin et al., 1962). Some of Guilford's (1959) work on divergent thinking also deals with stylistic restrictions in the generation and manipulation of ideas, which appear as much like personality consistencies as measures of "maximum performance" abilities (Cronbach, 1960).

In many cases, the objective nature of these tests makes it difficult to decide how to fake, since some look very much like ability tests and appear to have clear adaptive requirements that subjects should strive to achieve. Test properties, however, have been studied primarily in research contexts, where deliberate faking may have been minimal. Certain characteristics may change under other conditions. Available objective tests also tend to be unreliable, primarily because they have been deliberately kept short for use in large test batteries. Because of practice and order effects on some of the procedures, however, there is no guarantee that high reliabilities can be obtained simply by lengthening the tests.

Considerable attention has been given in recent years to certain stylistic dimensions in the performance of cognitive tasks (Witkin et al., 1954; Witkin et al., 1962; Gardner et al., 1959; Gardner, Jackson, & Messick, 1960). These personality dimensions have been conceptualized as cognitive styles, which represent a person's typical modes of perceiving, remembering, thinking, and problem-solving. Approaches to the measurement of these variables have routinely included objective procedures. Some examples of these dimensions are (1) *field-dependence-independence* – "an analytical, in contrast to a global, way of perceiving (which) entails a tendency to experience items as discrete from their backgrounds and reflects ability to overcome the influence of an embedding context" (Witkin et al., 1962; see also Kagan, Moss, & Sigel, 1963; Messick & Fritzky, 1963); (2) *leveling-sharpening* – a dimension where subjects at the leveling extreme tend to assimilate new material to an established framework, whereas sharpeners, at the other extreme, tend to contrast new material with the old and to maintain distinctions (Gardner et al., 1959); and (3) *category-width preferences*—a dimension of individual consistencies in modes of categorizing perceived similarities and differences, reflected in consistent preferences for broad or narrow categories in conceptualizing (Gardner et al., 1959; Gardner & Schoen, 1962; Pettigrew, 1958; Messick & Kogan, 1963a; Sloane, Gorlow, & Jackson, 1963).

Both the cognitive nature and the stylistic nature of these variables make them appear particularly relevant to the kinds of cognitive tasks performed in academic settings. Certain types of subject matter and certain problems or problem formulations might favor broad categorizers over narrow categorizers, for example, or levelers over sharpeners, and vice versa. This "vice versa" is extremely important: since it is unlikely that one end of such stylistic dimensions would prove uniformly more adaptive than the other, the relativity of their value should be recognized. (Incidentally, the possibility of such relativity of value might well be extended to other personality variables where the desirability of one end of the trait has usually been pre-judged. What conceptions would change, for example, if "flexibility vs. rigidity" had been called "confusion vs. control"?)

It is quite possible that we have already unwittingly included such stylistic variance in some measures of intellectual aptitude, such as the SAT, but if this is the case, the nature and direction of its operation should be specified and controlled. It is possible, for example, that the five-alternative multiple-choice form of quantitative aptitude items might favor subjects who prefer broad categories on category-width measures. Quick, rough approximations to the quantitative items might appropriately be judged by these subjects to be "close enough" to a given alternative, whereas "narrow range" subjects may require more time-consuming exact solutions before answering. Significant correlations between category preferences and quantitative aptitude tests have indeed been obtained and have been found to vary widely as a function of the spacing of alternatives on multiple-choice forms of the quantitative aptitude tests (Messick & Kogan, 1963b).

The Ethics of Selection

In considering personality measures of potential utility in the evaluative context of college performance, I have tried to give the impression that many measures are available but none is adequate when systematically evaluated against psychometric standards. In addition, I have tried to give some indication of the rapidly advancing technology that is evolving in personality measurement to support research efforts. In the relatively near future, this technology may produce personality measures that are acceptable by measurement and prediction standards, so that the question may soon arise in earnest as to the scope of their practical application. We have considered some of the scientific standards for deciding their appropriateness, but what about the ethical ones?

The choice of any particular personality measure for use, say, in college admission involves an implicit value judgment, which, at the least, should be made explicit in an educational policy that attempts to justify its use. One compelling justification for using personality measures in college selection would be to screen out extreme deviants. Colleges would be well advised, for example, to consider rejecting assaultive or suicidal psychotics, and some schools might wish to eliminate overt homosexuals. The use of personality measures for differentiating among normal subjects might also be justified in terms of empirical validity. After all, as long as there are many more candidates for admission than can be accepted, it seems better to make selections on the basis of valid measures than on the basis of chance. But is empirical validity enough? Validity for what? Certainly the role of the criterion in such an argument must be clearly specified.

The relevant domain of criterion performances should be outlined and attempts should be made to develop appropriate criterion measures. Since different criterion domains can be defined for different aspects of college success, selection might be oriented toward several of them simultaneously or toward only a few. Consider some of the possibilities: In *selection for academic performance,* criterion measures might include global grade-point averages, separate grades for different subject-matter fields, or standardized curriculum achievement examinations. In *selection for college environment,* criterion measures could be set up in terms of desired contributions to extra-curricular college life (such as football playing and newspaper editing) or in terms of balancing geographic, social class, sex, and, perhaps, temperament distributions in the student body. If the demands and pressures of the college environment and social structure have been studied, criterion standards might also be specified for selecting students with congenial needs that will fit well with (and hopefully have a higher probability of being satisfied by) the college environment (cf. Stern, 1962). We could also talk in terms of *selection for ultimate career satisfaction* and *selection for desirable personal characteristics* (cf. Davis, 1963)—or for desirable attitudes.

In each of these cases, it should be emphasized that potential predictor measures are not evaluated in terms of their empirical validity for *criterion behaviors* but rather in terms of their prediction of *criterion measures,* which, in turn, are presumed to reflect the criterion behaviors of interest. And these criterion measures should be evaluated against the same psychometric standards as any other measures. Not only should they be reliable, but also the nature of the attributes measured should be elucidated in a construct validity framework (Dunnette, 1963). Since each of these criterion measures may also contain some specific variance that is not particularly related to the criterion behaviors, one should also be concerned that an obtained validity coefficient reflects a correlation with relevant domain characteristics and not with irrelevant variance incidentally reflected in the putative criterion measure. Thus, the question of the *intrinsic validity* of the predictor and of the criterion measures should be broached, even if in practice many of the answers may seem presumptive ones (Gulliksen, 1950). In the last analysis, ultimate criteria are determined on rational grounds in any event (Thorndike, 1949). Should a reading comprehension test predict grades in gunner's mate school (Frederiksen, 1948)? Should a college that found docile, submissive students receiving higher grades in freshman courses select on this basis, or should they consider revising their grading system? Such decisions might become more difficult if the personality characteristics involved had more socially desirable labels.

Just as we have been concerned about predicting grades not as they are but as they should be (Frederiksen, 1954; Fishman, 1958), so too should we be concerned not only with predicting personality characteristics that are presently considered desirable for college students but also with deciding which characteristics, if any, should be considered desirable. It is possible, for example, that certain prepotent values, such as the desire for diversity, would override decisions to select students in terms of particular personal qualities. The very initiation of selection on any given personality variables might lead to conformity pressures toward the stereotype implied by the selected characteristics. Apart from the effects of the selection itself, such pressures to simulate desired personal qualities would probably decrease diversity in the college environment and in the personalities of the students. Wolfle (1960) and others have emphasized the value of diversity and even the value of uneven acquisition of skills within individuals as important contributors to the optimal development of talent. Restrictions upon diversity, however subtle, should therefore be undertaken cautiously.

I'd like to close metaphorically with a story of the lineage of King Arthur. At the end of the second book of *The Once and Future King*, T. H. White points out that Arthur's half-sister bore him a son, Modred, who was his ultimate downfall, that on the eve of the conception Arthur was a very young man drunk with the spoils of recent victory, that his half-sister was much older than he and active in the seduction, and that Arthur did not know that the woman was his sister. But it seems that "in tragedy, innocence is not enough." And in the use of personality measures in college admission, empirical validity is not enough.

REFERENCES

Bechtoldt, H. P. "Construct Validity: A Critique." *American Psychologist,* 1959, 14, 619-629.

Becker, W. C. "The Matching of Behavior Rating and Questionnaire Personality Factors." *Psychological Bulletin,* 1960, 57, 201-212.

Becker, W. C. "A Comparison of the Factor Structure and Other Properties of the 16 P. F. and the Guilford-Martin Personality Inventories. *Educational* and *Psychological Measurement,* 1961, 21, 393-404.

Bendig, A. W. "'Social desirability' and 'Anxiety' Variables in the IPAT Anxiety Scale." *Journal of Consulting Psychology,* 1959, 23, 377.

Berg, I. A. "Response Bias and Personality: the Deviation Hypothesis. *Journal of Psychology,* 1955, 40, 60-71.

Borislow, B. "The Edwards Personal Preference Schedule (EPPS) and Fakability." *Journal of Applied Psychology,* 1958, 42, 22-27.

Campbell, D. T."The Indirect Assessment of Social Attitudes." *Psychological Bulletin,* 1950, 47, 15-38.

Campbell, D. T. "A Typology of Tests, Projective and Otherwise." *Journal of Consulting Psychology,* 1957, 21, 207-210.

Campbell, D. T. "Recommendations for APA Test Standards Regarding Construct, Trait, or Discriminant Validity." *American Psychologist,* 1960, 15, 546-553.

Campbell, D. T., & Fiske, D. W. "Convergent and Discriminant Validation by the Multitrait-multimethod Matrix." *Psychological Bulletin,* 1959, 56, 81-105.

Cattell, R. B. *Personality and Motivation Structure and Measurement.* Yonkers-on-Hudson, N. Y.: World Book Co., 1957.

Comrey, A. "Factored Homogeneous Item Dimensions: A Strategy for Personality Research." In S. Messick & J. Ross (Eds.), *Measurement in Personality and Cognition.* New York: Wiley, 1962.

Corah, N. L., Feldman, M. J., Cohen, I. S., Gruen, W., Meadow, A., & Ringwall, E. A. "Social Desirability as a Variable in the Edwards Personal Preference Schedule." *Journal of Consulting Psychology,* 1958, 22, 70-72.

Couch, A., & Keniston, K. "Yeasayers and Naysayers: Agreeing Response Set as a Personality Variable." *Journal of Abnormal and Social Psychology,* 1960, 60, 151-174.

Cronbach, L. J. "Further Evidence on Response Sets and Test Design." *Educational and Psychological Measurement,* 1950, 10, 3-31.

Cronbach, L. J. "Review of Basic readings on the MMPI in Psychology and Medicine" (G. S. Welsh & W. G. Dahlstrom, Eds.), *Psychometrika,* 1958, 23, 385-386.

Cronbach, L. J. *Essentials of Psychological Testing.* New York: Harpers, 1960.

Cronbach, L. J. & Meehl, P. E. "Construct Validity in Psychological Tests." *Psychological Bulletin*, 1955, 52, 281-302.

Dahlstrom, W. G., & Welsh, G. S. *An MMPI Handbook*. Minneapolis: University of Minnesota Press, 1960.

Davis, J. A. "Desirable Characteristics of College Students: The Criterion Problem." Paper read at the APA symposium on "New Concepts and Devices in Measurement," 1963.

Dunnette, M. D. "A Note on *the* Criterion." *Journal of Applied Psychology*, 1963, 47, 251-254.

Ebel, R. L. "Must All Tests Be Valid?" *American Psychologist*, 1961, 16, 640-647.

Edwards, A. L. *The Social Desirability Variable in Personality Assessment and Research*. New York: Dryden, 1957.

Edwards, A. L., Diers, C. J., & Walker, J. N. "Response Sets and Factor Loadings on 61 Personality Scales." *Journal of Applied Psychology*, 1962, 46, 220-225.

Edwards, A. L., Wright, C. E., & Lunneborg, C. E. "A Note on 'Social Desirability as a Variable in the Edwards Personal Preference Schedule.'" *Journal of Consulting Psychology*, 1959, 23, 558.

Fishman, J. A. "Unsolved Criterion Problems in the Selection of College Students." *Harvard Educational Review*, 1958, 28, 340-349.

Frederiksen, N. "Statistical Study of the Achievement Testing Program in Gunner's Mates Schools." (Navpers 18079), 1948.

Frederiksen, N. "The Evaluation of Personal and Social Qualities." In *College Admissions*. New York: College Entrance Examination Board, 1954.

Frenkel-Brunswik, Else. "Mechanisms of Self Deception." *Journal of Social Psychology*, 1939, 10, 409-420.

Gardner, R. W., Holzman, P. S., Klein, G. S., Linton, H. B., & Spence, D. P. "Cognitive Control." *Psychological Issues*, 1959, 1, Monograph 4.

Gardner, R. W., Jackson, D. N., & Messick, S. "Personality Organization in Cognitive Controls and Intellectual Abilities." *Psychological Issues*, 1960, 2, Monograph 8.

Gardner, R. W., & Schoen, R. A. "Differentiation and Abstraction in Concept Formation." *Psychological Monographs*, 1962, 76, No. 41 (Whole No. 560).

Goffman, E. *The Presentation of Self in Everyday Life*. New York: Doubleday Anchor Books, 1959.

Guilford, J. P. *Personality*. New York: McGraw-Hill, 1959.

Guilford, J. P., & Zimmerman, W. S. "Fourteen Dimensions of Temperament." *Psychological Monographs*, 1956, 70, Whole No. 417.

Gulliksen, H. "Intrinsic Validity." *American Psychologist*, 1950, 5, 511-517.

Heilbrun, A. B., & Goodstein, L. D. "Relationships Between Personal and Social Desirability Sets and Performance on The Edwards Personal Preference Schedule." *Journal of Applied Psychology,* 1959, 43, 302-305.

Henrysson, S. "The Relation Between Factor Loadings and Biserial Correlations in Item Analysis." *Psychometrika,* 1962, 27, 419-424.

Jackson, D. N. "Stylistic Response Determinants in the California Psychological Inventory." *Educational and Psychological Measurement,* 1960, 20, 339-346.

Jackson, D. N., & Messick, S. "Content and Style in Personality Assessment." *Psychological Bulletin,* 1958, 55, 243-252.

Jackson, D. N., & Messick, S. "Acquiescence and Desirability as Response Determinants on the MMPI." *Educational and Psychological Measurement,* 1961, 21, 771-790.

Jackson, D. N., & Messick, S. "Response Styles on the MMPI: Comparison of Clinical and Normal Samples." *Journal of Abnormal and Social Psychology,* 1962, 65, 285-299. (a)

Jackson, D. N., & Messick, S. "Response Styles and the Assessment of Psycho-Pathology." In S. Messick & J. Ross (Eds.), *Measurement in Personality and Cognition.* New York: Wiley, 1962. (b)

Kagan, J., Moss, H. A., & Sigel, I. E. "The Psychological Significance of Styles of Conceptualization." In J. C. Wright & J. Kagan (Eds)," Basic Cognitive Processes in Children." *Monographs of the Society for Research in Child Development,* 1963, 28, No. 2, 73-112.

LaPointe, R. E., & Auclair, G. A. "The Use of Social Desirability in Forced-choice Methodology." *American Psychologist,* 1961, 16, 446 (Abstract).

Lennon, R. "Assumptions Underlying the Use of Content Validity." *Educational and Psychological Measurement,* 1956, 16, 294-304.

Loevinger, Jane. "Some Principles of Personality Measurement." *Educational and Psychological Measurement,* 1955, 15, 3-17.

Loevinger, Jane. "Objective Tests as Instruments of Psychological Theory." *Psychological Reports,* 1957, 3, 635-694.

Loevinger, Jane, Gleser, Goldine, & Dubois, P. H. "Maximizing the Discriminating Power of a Multiple-score Test." *Psychometrika,* 1953, 18, 309-317.

Meehl, P. E., & Hathaway, S. R. "The K Factor as a Suppressor Variable in the MMPI." *Journal of Applied Psychology,* 1946, 30, 525-564.

Messick, S. "Dimensions of Social Desirability." *Journal of Consulting Psychology,* 1960, 24, 279-287.

Messick, S. "Response Style and Content Measures From Personality Inventories." *Educational and Psychological Measurement,* 1962, 22, 41-56.

Messick, S., & Fritzky, F. J. "Dimensions of Analytic Attitude in Cognition and Personality." *Journal of Personality,* 1963, 31, 346-370.

Messick, S., & Jackson, D. N. "Acquiescence and the Factorial Interpretation of the MMPI." *Psychological Bulletin,* 1961, 58, 299-304.

Messick, S., & Kogan, N. "Differentiation and Compartmentalization in Object-sorting Measures of Categorizing Style." *Perceptual and Motor Skills*, 1963, 16, 47-51. (a)

Messick, S., & Kogan, N. "Category Width and Quantitative Aptitude." Princeton, N. J.: Educational Testing Service, Research Bulletin, 1963. (b)

Norman, W. T. "Toward an Adequate Taxonomy of Personality Attributes: Replicated Factor Structure in Peer Nomination Personality Ratings." *Journal of Abnormal and Social Psychology*, 1963, 66, 574-583. (a)

Norman, W. T. "Personality Measurement, Faking, and Detection: An Assessment Method for Use in Personal Selection." *Journal of Applied Psychology*, 1963, 47, 225-241. (b)

Peabody, D. "Two Components in Bipolar Scales: Direction and Extremeness." *Psychological Review*, 1962, 69, 65-73.

Pettigrew, T. F. "The Measurement and Correlates of Category Width as a Cognitive Variable." *Journal of Personality*, 1958, 26, 532-544.

Rosen, E. "Self-appraisal, Personal Desirability, and Perceived Social Desirability of Personality Traits." *Journal of Abnormal and Social Psychology*, 1956, 52, 151-158.

Sechrest, L. B., & Jackson, D. N. "Deviant Response Tendencies: Their Measurement and Interpretation." *Educational and Psychological Measurement*, 1963, 23, 35-53.

Sloane, H. N., Gorlow, L., & Jackson, D. N. "Cognitive Styles in Equivalence Range." *Perceptual and Motor Skills*, 1963, 16, 389-404.

"Standards of Ethical Behavior for Psychologists." *American Psychologist*, 1963, 18, No. 1, 56-60.

Stern, G. G. "The Measurement of Psychological Characteristics of Students and Learning Environments." In S. Messick & J. Ross (Eds.), *Measurement in Personality and Cognition*. New York: Wiley, 1962.

Stricker, L. J. "A Review of the Edwards Personal Preference Schedule." In O. K. Buros (Ed.), *The Sixth Mental Measurements Yearbook*. Highland Park, N. J.: Gryphon Press, 1963 (in press).

Thorndike, R. L. *Personnel Selection*. New York: Wiley, 1949.

Thurstone, L. L. *A Factorial Study of Perception*. Chicago: Univer. Chicago Press, Psychometric Monograph No. 4, 1944.

Tupes, E. C. "Relationships Between Behavior Trait Ratings by Peers and Later Officer Performance of USAF Officer Candidate School Graduates." USAF PTRC tech. Note, 1957, No. 57-125.

Tupes, E. C., & Christal, R. E. "Stability of Personality Trait Rating Factors Obtained under Diverse Conditions." USAF WADC tech. Note, 1958, No. 58-61.

Tupes, E. C., & Christal, R. E. "Recurrent Personality Factors Based on Trait Ratings." USAF ASD tech. Rep., 1961, No. 61-97.

Wahler, H. J. "Response Styles in Clinical and Nonclinical Groups." *Journal of Consulting Psychology,* 1961, 25, 533-539.

Welsh, G. S., & Dahlstrom, W. G. (Eds.). *Basic Readings on the MMPI in Psychology and Medicine.* Minneapolis: University of Minnesota Press, 1956.

Wherry, R. J., & Winer, B. J. "A Method for Factoring Large Numbers of Items." *Psychometrika,* 1953, 18, 161-179.

Whisler, T. L., & Harper, Shirley F. (Eds.), *Performance Appraisal.* New York: Holt, Rinehart, & Winston, 1962.

White, T. H. *The Once and Future King.* New York: Putnam, 1958.

Wiggins, J. S. "Interrelations Among MMPI Measures of Dissimulation under Standard and Social Desirability Instructions." *Journal of Consulting Psychology*, 1959, 23, 419-427.

Wiggins, J. S. "Strategic, Method, and Stylistic Variance in the MMPI." *Psychological Bulletin*, 1962, 59, 224-242.

Witkin, H. A., Lewis, H. B., Hertzman, M., Machover, K., Meissner, P. B., & Wapner, S. *Personality Through Perception.* New York: Harpers, 1954.

Witkin, H. A., Dyk, R. B., Faterson, H. F., Goodenough, D. R., & Karp, S. A. *Psychological Differentiation.* New York: John Wiley & Sons, Inc., 1962.

Wolfle, D. "Diversity of Talent." *American Psychologist,* 1960, 15, 535-545.

ROBERT L. EBEL,
*School for
Advanced Studies,
College of
Education,
Michigan State University*

The Social Consequences of Educational Testing

I have an uneasy feeling that some of the things that will be said in this talk on the social consequences of educational testing may be regarded as somewhat controversial. Let me try to begin, therefore, with some statements on which we may all be able to agree.

Popularity and Criticism

Tests have been used increasingly in recent years to make educational assessments. The reasons for this are not hard to discover. Educational tests of aptitude and achievement greatly improve the precision, objectivity and efficiency of the observations on which educational assessments rest. Tests are not alternatives to observations. At best they represent no more than refined and systematized processes of observation.

But the increasing use of tests has been accompanied by an increasing flow of critical comment. Again the reasons are easy to see. Tests vary in quality. None is perfect and some may be quite imperfect. Test scores are sometimes misused. And even if they were flawless and used with the greatest skill, they would probably still be unpopular among those who have reason to fear an impartial assessment of some of their competencies.

Many of the popular articles critical of educational testing that have appeared in recent years do not reflect a very adequate understanding of educational testing, or a very thoughtful, unbiased consideration of its social consequences. Most of them are obvious potboilers for their authors, and sensational reader-bait in the eyes of the editors of the journals in which they appear. The writers of some of these articles have paid courteous visits to our offices. They have listened respectfully to our recitals of fact and opinion. They have drunk coffee with us and then taken their leave, presumably to reflect on what they have been told, but in any event, to write. What appears in print often seems to be only an elaboration and documentation of their initial prejudices and preconceptions, supported by atypical anecdotes and purposefully selected quotations. Educational testing has not fared very well in their hands.

Among the charges of malfeasance and misfeasance that these critics have leveled against the test makers there is one of nonfeasance. Specifically, we are charged with having shown lack of proper concern for the social consequences of our educational testing. These harmful consequences, they have suggested, may be numerous and serious. The more radical among them imply that, because of what they suspect about the serious social consequences of educational testing, the whole testing movement ought to be suppressed. The more moderate critics claim that they do not know much about these social consequences. But they also suggest that the test makers don't either, and that it is the test makers who ought to be doing substantial research to find out.

The Role of Research

If we were forced to choose between the two alternatives offered by the critics, either the suppression of educational testing or extensive research on its social consequences, we probably would choose the latter without much hesitation. But it is by no means clear that what testing needs most at this point is a large program of research on its social consequences. Let me elaborate.

Research can be extremely useful, but it is far from being a sure-fire process for finding the answers to any kind of a question, particularly a social question, that perplexes us. Nor is research the only source of reliable knowledge. In the social sciences, at least, most of what we know for sure has not come out of formal research projects. It has come instead from the integration of a very large number of more or less incidental observations and accounts of human behavior in natural, rather than experimental, situations. There are good reasons why research on human behavior tends to be difficult, and often unproductive, but that is a story we cannot go into now.

For present purposes, only two points need to be mentioned. The first is that the scarcity of formal research on the social consequences of educational testing should not be taken to mean that there is no reliable knowledge about those consequences, or that those engaged in educational testing have been callously indifferent to its social consequences. The second is that scientific research on human behavior may require commitment to values that are in basic conflict with our democratic concerns for individual welfare. If boys and girls are used as carefully controlled experimental subjects in tough-minded research on social issues that really matter, not all of them will benefit, and some may be disadvantaged seriously. Our society is not yet ready, and perhaps should never become ready to acquiesce in that kind of scientific research.

Harmful Consequences

Before proceeding further, let us mention specifically a few of the harmful things that critics have suggested educational testing may do:

- It may place an indelible stamp of intellectual status—superior, mediocre or inferior—on a child, and thus predetermine his social status as an adult, and possibly also do irreparable harm to his self-esteem and his educational motivation.
- It may lead to a narrow conception of ability, encourage pursuit of this single goal, and thus tend to reduce the diversity of talent available to society.
- It may place the testers in a position to control education and determine the destinies of individual human beings, while, incidentally, making the testers themselves rich in the process.

- It may encourage impersonal, inflexible, mechanistic processes of evaluation and determination, so that essential human freedoms are limited or lost altogether.

These are four of the most frequent and serious tentative indictments. There have been, of course, many other suggestions of possible harmful social consequences of educational testing. It may emphasize individual competition and success, rather than social cooperation, and thus conflict with the cultivation of democratic ideals of human equality. It may foster conformity rather than creativity. It may involve cultural bias. It may neglect important intangibles. It may, particularly in the case of personality testing, involve unwarranted and offensive invasions of privacy. It may do serious injustice in particular individual cases. It may reward specious test-taking skill, or penalize the lack of it.

If time and our supply of ideas permitted, it would be well for us to consider all of these possibilities. But since they do not, perhaps the demands of the topic may be reasonably well met if we limit attention to the first four items mentioned as possibly harmful consequences of educational testing, namely:

- permanent status determination
- limited conceptions of ability
- domination by the testers
- mechanistic decision making

At this point in the presentation, a major choice must be made. Shall we explore the foundations for these apprehensions and attempt to dispel them? Shall we, in other words, attempt to refute the allegations of harmful social consequences of educational testing? Clearly most of these social dangers can be, and probably have been, exaggerated. Little solid evidence exists to justify the fears that have been expressed with such apparent concern.

Or shall we assume that the concerns which have been expressed are not wholly fanciful? Shall we, therefore, set as our task the discovery and delineation of things that might be done by those who make and use tests to limit the causes for concern. On reflection it seemed that for one speaking to a group of specialists in educational testing, the second course of action was clearly the more reasonable, and would be likely to be the more useful. So that is the course that has been chosen.

PERMANENT STATUS DETERMINATION

Consider first, then, the danger that educational testing may place an indelible stamp of inferiority on a child, ruin his self-esteem and educational motivation, and determine his social status as an adult. The kind of educational testing most likely to have these consequences would involve tests purporting to measure a person's permanent general capacity for learning. These are the intelligence tests, and the presumed measures of general capacity for learning they provide are popularly known as IQ's.

Most of us here assembled are well aware of the fact that there is no direct, unequivocal means for measuring permanent general capacity for learning. It is not even clear to many of us that, in the state of our current understanding of mental functions and the learning process, any precise and useful meaning can be given to the concept of "permanent general capacity for learning." We know that all intelligence tests now available are direct measures only of achievement in learning, including learning how to learn, and that inferences from scores on those tests to some native capacity for learning are fraught with many hazards and uncertainties.

But many people who are interested in education do not know this. Many of them believe that native intelligence has been clearly indentified and is well understood by expert psychologists. They believe that a person's IQ is one of his basic, permanent attributes, and that any good intelligence test will measure it with a high degree of precision. They do not regard an IQ simply as another test score, a score that may vary considerably depending on the particular test used and the particular time when the person was tested.

Whether or not a person's learning is significantly influenced by his predetermined capacity for learning, there is no denying the obvious fact that individual achievements in learning exhibit considerable consistency over time and across tasks. The superior elementary school pupil may become a mediocre secondary school pupil and an inferior college student, but the odds are against it. Early promise is not always fulfilled, but it is more often than not. The A student in mathematics is a better bet than the C student to be an A student in English literature as well, or in social psychology.

On the other hand, early promise is not always followed by late fulfillment. Ordinary students do blossom sometimes into outstanding scholars. And special talents can be cultivated. There is enough variety in the work of the world so that almost anyone can discover some line of endeavor in which he can develop more skill than most of his fellow men.

In a free society that claims to recognize the dignity and worth of every individual, it is better to emphasize the opportunity for choice and the importance of effort than to stress genetic determinism of status and success. It is better to emphasize the diversity of talents and tasks than to stress general excellence or inferiority. It is important to recognize and to reinforce what John Gardner has called "the principle of multiple chances," not only across time but also across tasks.

The concept of fixed general intelligence, or capacity for learning, is a hypothetical concept. At this stage in the development of our understanding of human learning, it is not a necessary hypothesis. Socially, it is not now a useful hypothesis. One of the important things test specialists can do to improve the social consequences of educational testing is to discredit the popular conception of the IQ. Wilhelm Stern, the German psychologist who suggested the concept originally, saw how it was being overgeneralized and charged one of his students coming to America to "kill the IQ." Perhaps we would be well advised, even at this late date, to renew our efforts to carry out his wishes.

Recent emphasis on the early identification of academic talent involves similar risks of oversimplifying the concept of talent and overemphasizing its predetermined components. If we think of talent mainly as something that is genetically given, we will run our schools quite differently than if we think of it mainly as something that can be educationally developed.

If human experience, or that specialized branch of human experience we call scientific research, should ever make it quite clear that differences among men in achievement are largely due to genetically determined differences in talent, then we ought to accept the finding and restructure our society and social customs in accord with it. But that is by no means clear yet, and the structure and customs of our society are not consistent with such a basic assumption. For the present, it will be more consistent with the facts as we know them, and more constructive for the society in which we live, to think of talent not as a natural resource like gold or uranium to be discovered, extracted and refined, but as a synthetic product like fiberglass or D.D.T.—something that, with skill, effort and luck, can be created and produced out of generally available raw materials to suit our particular needs or fancies.

This means, among other things, that we should judge the value of the tests we use not in terms of how accurately they enable us to *predict* later achievement, but rather in terms of how much help they give us to *increase* achievement by motivating and directing the efforts of students and teachers. From this point of view, those concerned with professional education who have resisted schemes for very long-range predictions of aptitude for, or success in, their professions have acted wisely. Not only is there likely to be much more of dangerous error than of useful truth in such long-range predictions, but also there is implicit in the whole enterprise a deterministic conception of achievement that is not wholly consistent with the educational facts as we know them, and with the basic assumptions of a democratic, free society.

Whenever I try to point out that prediction is not the exclusive, nor even the principal purpose of educational measurement, some of my best and most intelligent friends demur firmly, or smile politely to communicate that they will never accept such heretical nonsense. When I imply that they use the term "prediction" too loosely, they reply that I conceive it too narrowly. Let me try once more to achieve a meeting of the minds.

I agree that prediction has to do with the future, and that the future ought to be of greater concern to us than the past. I agree, too, that a measurement must be related to some other measurements in order to be useful, and that these relationships provide the basis for, and are tested by, predictions. But these relationships also provide a basis, in many educational endeavors, for managing outcomes—for making happen what we want to happen. And I cannot agree that precision in language or clarity of thought is well served by referring to this process of controlling outcomes as just another instance of prediction. The etymology and common usage of the word "prediction" imply to me the process of foretelling, not of controlling.

The direct, exclusive, immediate purpose of measurement is always description, not either prediction or control. If we know with reasonable accuracy how things now stand (descriptions), and if we also know with reasonable accuracy what leads to what (functional relations), we are in a position to foretell what will happen if we keep hands off (prediction) or to manipulate the variables we can get our hands on to make happen what we want to happen (control). Of course, our powers of control are often limited and uncertain, just as our powers of prediction are. But I have not been able to see what useful purpose is served by referring to both the hands-off and the hands-on operations as prediction, as if there were no important difference between them. It is in the light of these semantic considerations that I suggest that tests should be used less as bases for prediction of achievement, and more as means to increase achievement. I think there is a difference, and that it is important educationally.

LIMITED CONCEPTIONS OF ABILITY

Consider next the danger that a single widely used test or test battery for selective admission or scholarship awards may foster an undesirably narrow conception of ability and thus tend to reduce diversity in the talents available to a school or to society.

Here again, it seems, the danger is not wholly imaginary. Basic as verbal and quantitative skills are to many phases of educational achievement, they do not encompass all phases of achievement. The application of a common yardstick of aptitude or achievement to all pupils is operationally much simpler than the use of a diversity of yardsticks, designed to measure different aspects of achievement. But overemphasis on a common test could lead educators to neglect those students whose special talents lie outside the common core.

Those who manage programs for the testing of scholastic aptitude always insist, and properly so, that scores on these tests should not be the sole consideration when decisions are made on admission or the award of scholarships. But the question of whether the testing itself should not be varied from person to person remains. The use of optional tests of achievement permits some variation. Perhaps the range of available options should be made much wider than it is at present to accommodate greater diversity of talents.

The problem of encouraging the development of various kinds of ability is, of course, much broader than the problem of testing. Widespread commitment to general education, with the requirement that all students study identical courses for a substantial part of their programs, may be a much greater deterrent of specialized diversity in the educational product. Perhaps these requirements should be restudied too.

DOMINATION BY THE TESTERS

What of the concern that the growth of educational testing may increase the influence of the test makers until they are in a position to control educational curricula and determine the destinies of students?

Those who know well how tests are made and used in American education know that the tests more often lag than lead curricular change, and that while tests may affect particular episodes in a student's experience, they can hardly ever be said to determine a student's destiny. American education is, after all, a manifold, decentralized, loosely organized enterprise. Whether it restricts student freedom too much or too little is a subject for lively debate. But it does not even come close to determining any student's destiny, not nearly as close as the examination systems in some other countries, ancient and modern.

But test makers have, I fear, sometimes given the general public reason to fear that we may be up to no good. I refer to our sometime reluctance to take the layman fully into our confidence, to share fully with him all our information about his test scores, the tests from which they were derived, and our interpretations of what they mean.

Secrecy concerning educational tests and test scores has been justified on several grounds. One is that the information is simply too complex for untrained minds to grasp. Now it is true that some pretty elaborate theories can be built around our testing processes. It is also true that we can perform some very fancy statistical manipulations with the scores they yield. But the essential information revealed by the scores on most educational tests is not particularly complex. If we understand it ourselves, we can communicate it clearly to most laymen without serious difficulty. To be quite candid, we are not all that much brighter than they are, much as we may sometimes need the reassurance of thinking so.

Another justification for secrecy is that laymen will misuse test scores. Mothers may compare scores over the back fences. The one whose child scores high spreads the word around. The one whose child scores low may keep the secret, but seek other grounds for urging changes in the teaching staff or in the educational program. Scores of limited meaning may be treated with undue respect and used to repair or to injure the student's self-esteem rather than to contribute to his learning.

Again it is true that test scores can be misused. They have been in the past and they will be in the future. But does this justify secrecy? Can we minimize abuses due to ignorance by withholding knowledge? We do not flatter our fellow citizens when we tell them, in effect, that they are too ignorant, or too lacking in character to be trusted with the knowledge of their children, or of themselves, that we possess.

Seldom acknowledged, but very persuasive as a practical reason for secrecy regarding test scores, is that it spares those who use the scores from having to explain and justify the decisions they make. Preference is not, and should not, always be given to the person whose test score is the higher. But if score information is withheld, the disappointed applicant will assume that it was because of his low score, not because of some other factor. He will not trouble officials with demands for justification of a decision that, in some cases, might be hard to justify. But all things considered, more is likely to be gained in the long run by revealing the objective evidence used in reaching a decision. Should the other, subjective considerations prove too difficult to justify, perhaps they ought not to be used as part of the basis for decision.

If specialists in educational measurement want to be properly understood and trusted by the public they serve, they will do well to shun secrecy and to share with the public as much as it is interested in knowing about the methods they use, the knowledge they gain, and the interpretations they make. This is clearly the trend of opinion in examining boards and public education authorities. Let us do what we can to reinforce the trend. Whatever mental measurements are so esoteric or so dangerous socially that they must be shrouded in secrecy probably should not be made in the first place.

The testers do not control education or the destinies of individual students. By the avoidance of mystery and secrecy, they can help to create better public understanding and support.

MECHANISTIC DECISION MAKING

Finally, let us consider briefly the possibility that testing may encourage mechanical decision making, at the expense of essential human freedoms of choice and action.

Those who work with mental tests often say that the purpose of all measurement is prediction. They use regression equations to predict grade point averages, or contingency tables to predict the chances of various degrees of success. Their procedures may seem to imply not only that human behavior is part of a deterministic system in which the number of relevant variables is manageably small, but also that the proper goals of human behavior are clearly known and universally accepted.

In these circumstances, there is some danger that we may forget our own inadequacies and attempt to play God with the lives of other human beings. We may find it convenient to overlook the gross inaccuracies that plague our measurements, and the great uncertainties that bedevil our predictions. Betrayed by overconfidence in our own wisdom and virtue, we may project our particular value systems into a pattern of ideal behavior for all men.

If these limitations on our ability to mould human behavior and to direct its development did not exist, we would need to face the issue debated by B. F. Skinner and Carl Rogers before the American Psychological Association some years ago. Shall our knowledge of human behavior be used to design an ideal culture and condition individuals to live happily in it at whatever necessary cost to their own freedom of choice and action?

But the aforementioned limitations do exist. If we ignore them and undertake to manage the lives of others so that those others will qualify as worthy citizens in our own particular vision of utopia, we do justify the concern that one harmful social consequence of educational testing may be mechanistic decision making and the loss of essential human freedoms.

A large proportion of the decisions affecting the welfare and destiny of a person must be made in the midst of overwhelming uncertainties concerning the outcomes to be desired and the best means of achieving such outcomes. That many mistakes will be made seems inevitable. One of the cornerstones of a free society is the belief that in most cases it is better for the person most concerned to make the decision, right or wrong, and to take the responsibility for its consequences, good or bad.

The implications of this for educational testing are clear. Tests should be used as little as possible to *impose* decisions and courses of action on others. They should be used as much as possible to provide a sounder basis of *choice* in individual decision making. Tests can be used and ought to be used to support rather than to limit human freedom and responsibility.

Conclusion

In summary, we have suggested here today that those who make and use educational tests might do four things to alleviate public concerns over their possibly adverse social consequences:

1. We could emphasize the use of tests to improve status, and de-emphasize their use to determine status.
2. We could broaden the base of achievements tested to recognize and develop the wide variety of talents needed in our society.
3. We could share openly with the persons most directly concerned all that tests have revealed to us about their abilities and prospects.
4. We could decrease the use of tests to impose decisions on others, and instead increase their use as a basis for better personal decision making.

When Paul Dressel read a draft of this paper, he chided me gently on what he considered to be a serious omission. I had failed to discuss the social consequences of *not* testing. What are some of these consequences?

If the use of educational tests were abandoned, the distinctions between competence and incompetence would become more difficult to discern. Dr. Nathan Womack, former president of the National Board of Medical Examiners, has pointed out that only to the degree to which educational institutions can define what they mean by competence, and determine the extent to which it has been achieved, can they discharge their obligation to deliver competence to the society they serve.

If the use of educational tests were abandoned, the encouragement and reward of individual efforts to learn would be made more difficult. Excellence in programs of education would become less tangible as a goal and less demonstrable as an attainment. Educational opportunities would be extended less on the basis of aptitude and merit and more on the basis of ancestry and influence, social class barriers would become less permeable. Decisions on important issues of curriculum and method would be made less on the basis of solid evidence and more on the basis of prejudice or caprice.

These are some of the social consequences of *not* testing. In our judgment, they are potentially far more harmful than any possible adverse consequences of testing. But it is also our judgment, and has been the theme of this paper, that we can do much to minimize even these possibilities of harmful consequences. Let us, then, use educational tests for the powerful tools they are with energy and skill, but also with wisdom and care.

Participants

1963 Invitational Conference on Testing Problems

ABBOTT, MURIEL M., Upper Montclair, New Jersey
ABRAHAM, ANSLEY A., Florida Agricultural and Mechanical University
ADKINS, DOROTHY C., University of North Carolina
AHRENS, DOLORES F., Educational Testing Service
ALBRIGHT, FRANK S., West Orange (New Jersey) Public Schools
ALEXANDER, JEAN, Educational Testing Service
ALMAN, JOHN E., Boston University
ALT, PAULINE M., Central Connecticut State College
ALTSCHUL, KENNETH, Queens College
ANASTASI, ANNE, Fordham University
ANDERSON, G. ERNEST, JR., Newton Public Schools, Newtonville, Massachusetts
ANDERSON, HOWARD R., Houghton Mifflin Company
ANDERSON, ROSE G., New York City
ANDERSON, ROY N., North Carolina State College
ANDERSON, SCARVIA B., Educational Testing Service
ANDREWS, T. G., University of Maryland
ANGOFF, WILLIAM H., Educational Testing Service
ANNE, SISTER ELIZABETH, C.S.N., Saint Mary's School, Peekskill, New York
ANTES, CHARLOTTE, Educational Testing Service
ANTHONY, BROTHER E., F.S.C., LaSalle College
ARONOW, MIRIAM S., New York City Board of Education
ASH, PHILIP, Inland Steel Company
ASHCRAFT, KENNETH B., Colorado State Department of Education
ATKINS, WILLIAM H., Yeshiva University
AUKES, LEWIS E., University of Illinois
AUSTIN, NEALE W., Educational Testing Service
AYRER, JAMES, New Jersey State Department of Civil Service
BALL, ALBERT T., Monmouth College
BANNON, CHARLES J., Department of Education, Waterbury, Connecticut
BANNON, MARION K., Department of Education, Waterbury, Connecticut
BAPTISTA, SISTER MARIE, Boorady Reading Center, Dunkirk, New York
BARBARA, SISTER, S.C., College of Mount St. Joseph
BARBULA, MARVIN, University of Michigan
BARNES, PAUL J., Harcourt, Brace and World, Inc.
BARNETTE, W. L., JR., State University of New York at Buffalo
BARNETTE, EDMUND L., South Carolina State Department of Education

BARRETT, DOROTHY M., Hunter College

BARRETT, JUANITA, American and Foreign Teachers' Agency, New York City

BARRON, ARLEEN S., Educational Testing Service

BARTNIK, ROBERT V., Educational Testing Service

BATES, MARGARET A., Educational Testing Service

BEALS, ERNEST W., New Hampshire State Department of Education

BECHTOLD, DONALD, The Catholic University of America

BECK, JARRETTE A., U. S. Army Southeastern Signal School

BECKINGHAM, KATHLEEN R., University of New Hampshire

BENNETT, GEORGE K., The Psychological Corporation

BENNETT, MARJORIE G., Bronxville, New York

BENNINGTON, NEVILLE L., New York State Department of Education

BENSON, ARTHUR L., Educational Testing Service

BENSON, LOREN, Hopkins (Minnesota) High School

BENTZ, V. JON, Sears, Roebuck and Company, Chicago, Illinois

BERDIE, RALPH F., University of Minnesota

BERG, JOEL, King Philip Junior High School, West Hartford, Connecticut

BERGER, BERNARD, New York City Department of Personnel

BERGESEN, B. E., JR., Personnel Press, Inc., Princeton, New Jersey

BERGSTEIN, HARRY B., Huntington (New York) Public Schools

BERNE, ELLIS J., U. S. Department of Health, Education and Welfare

BEST, PHILLIP J., Educational Testing Service

BINGHAM, WILLIAM C., Rutgers, The State University

BIRNEY, ROBERT C., Amherst Colleg

BLACK, HILLEL, Saturday Evening Post

BLANCHARD, CARROLL M., U. S. Army Signal Center and School

BLANCHARD, D. D., Wilton (Connecticut) High School

BLEEKE, DONALD E., College Reading-Study and Counseling Center, Westfield, New Jersey

BLIGH, HAROLD F., Harcourt, Brace and World, Inc.

BLOOM, BENJAMIN S., University of Chicago

BLOOMER, RICHARD H., University of Connecticut

BLUM, STUART H., Hofstra University

BOLLENBACHER, JOAN, Cincinnati (Ohio) Public Schools

BONDARUK, JOHN, National Security Agency

BOUCHARD, ROBERT, New York State Department of Civil Service

BOUTWELL, WILLIAM DOW, Scholastic Teacher

BOWES, EDWARD W., University of California

BOWMAN, HOWARD A., Los Angeles (California) City Schools

BOYD, JOSEPH L., JR., Educational Testing Service
BRACA, SUSAN E., Oceanside (New York) Senior High School
BRADEN, BILLY, Kentucky State Department of Education
BRANDT, HYMAN, Programmed Instruction for Industry, New York City
BRANSFORD, THOMAS L., New York State Civil Service Department
BREAM, LOIS GOULD, Cheltenham Township (Pennsylvania) High School
BRICKELL, HELEN, Bronxville (New York) Senior School
BRISTOW, WILLIAM H., New York City Board of Education
BROOKS, RICHARD B., Longwood College
BROWN, FRANK, Melbourne (Florida) High School
BROWN, F. MA[illegible]IN, Fountain Valley School, Colorado Springs, Colorado
BROWN, FREDERICK S., Great Neck (New York) Public Schools
BRUNER, JEROME S., Harvard University
BRYAN, GLENN L., Office of Naval Research, Washington, D. C.
BRYAN, J. NED, JR., United States Office of Education
BRYAN, MIRIAM M., Educational Testing Service
BUDDEKE, SISTER RITA, Catholic University of America
BUNDERSON, C. VICTOR, Educational Testing Service
BUNDERSON, EILEEN D., Educational Testing Service
BURDOCK, E. I., The City University of New York
BURKE, JAMES M., Norwalk (Connecticut) Public Schools
BURKE, PAUL J., Bell Telephone Laboratories, Inc.
BURNHAM, PAUL S., Yale University
BUROS, OSCAR K., Rutgers, The State University
BURR, WILLIAM L., Harcourt, Brace and World, Inc.
BUSHNELL, DON D., System Development Corporation
CAFFREY, JOHN, System Development Corporation
CAHEN, LEONARD S., Stanford University
CAMPBELL, JOEL T., Educational Testing Service
CAPPS, MARIAN P., Virginia State College, Norfolk Division
CAPPUCCINO, JOHN, New Jersey State Department of Civil Service
CARNEGIE, ELIZABETH, Nursing Outlook, New York City
CARSON, JOHN, Pennsbury High School, Yardley, Pennsylvania
CARSTATER, EUGENE D., Bureau of Naval Personnel, Washington, D. C.
CARSTATER, MARIE H., Falls Church (Virginia) High School
CASS, JAMES, Saturday Review Education Supplement
CASSERLY, PATRICIA, Educational Testing Service
CAWLEY, JAMES F., Vermont State Department of Education
CHAMBERS, WILLIAM M., University of Kentucky
CHANDLER, THOMAS E., United States Army Southeastern Signal School

CHAPMAN, GLORIA, New York University
CHAPPELL, BARTLETT E. S., New York Military Academy
CHAUNCEY, HENRY, Educational Testing Service
CHESTNUT, CORA MAE, Rochester (New York) City School District
CHUGH, H. K. L., National Institute of Education, New Delhi, India
CIERI, VINCENT P., United States Army Signal Center and School
CLARK, PHILIP I., California Test Bureau, Scarsdale, New York
CLEARY, J. ROBERT, Educational Testing Service
CLEMANS, WILLIAM V., Science Research Associates
CLENDENEN, DOROTHY M., The Psychological Corporation
COFFMAN, WILLIAM E., Educational Testing Service
COLEMAN, ELIZABETH R., Clarkstown Junior High School, West Nyack, New York
COLLEEN, SISTER MARY, S.S.N.D., Archdiocese of New Orleans
COLVER, ROBERT M., Duke University
CONKLIN, NANCY, University of Rochester
CONLAN, GERTRUDE C., Educational Testing Service
CONLON, JOHN, Northern Valley Regional High School, Demarest, New Jersey
CONNELL, ELLEN F., Erie (Pennsylvania) School District
CONNOLLY, JOHN A., Educational Testing Service
COOPERMAN, IRENE G., Veterans Administration
COPE, WILLIAM E., JR., United Negro College Fund
COPELAND, HERMAN A., Pennsylvania State Civil Service Commission
COREY, STEPHEN M., Teachers College, Columbia University
CORNOG, WILLIAM H., New Trier Township (Illinois) High School
CORY, CHARLES H., Philadelphia Personnel Department
COUTANT, MADELINE F., Laurens (New York) Central School
COWLES, JOHN T., University of Pittsburgh
COX, HENRY M., The University of Nebraska
CRAMER, AILEEN L., Educational Testing Service
CRAVEN, ETHEL C., Veterans Administration
CREAGER, JOHN A., National Academy of Sciences
CROCKETT, DAVID S., American College Testing Program
CROSS, K. PATRICIA, Educational Testing Service
CROSS, ORRIN H., West Virginia University
CUMMINGS, MARY B., Boston (Massachusetts) Public Schools
CURETON, EDWARD E., University of Tennessee
CURETON, LOUISE W., University of Tennessee
CURRY, JOHN C., Mountain High School, West Orange, New Jersey

CURRY, ROBERT P., Cincinnati (Ohio) Public Schools
CUSUMANO, GLORIA, New York University
CUYLER, REVEREND CORNELIUS M., S.S., St. Charles College
CYNAMON, MANUEL, Brooklyn College
DALY, ALICE T., New York State Department of Education
D'AMOUR, O'NEIL C., National Catholic Educational Association
DANA, RICHARD H., West Virginia University
D'ARCY, EDWARD, New Jersey State Department of Civil Service
DAVIDOFF, MELVIN D., United States Civil Service Commission
DAVIDSON, HELEN H., The City College of New York
DAVIS, AILEEN H., Public Schools of the District of Columbia
DAVIS, JUNIUS A., Educational Testing Service
DE BRULER, RALPH M., Edmonds (Washington) School District 15
DE BURLO, C. RUSSELL, JR., Educational Testing Service
DENISON, VIOLET, Teachers College, Columbia University
DESMOND, RICHARD S., Paterson State College
DIAMOND, LORRAINE K., Teachers College, Columbia University
DICKSON, GWEN S., Peace Corps
DIEDERICH, PAUL B., Educational Testing Service
DIEHL, MARY JANE, Educational Testing Service
DIEHL, MICHAEL, Harbourton, New Jersey
DIGGS, FRANKLIN B., New York City Department of Personnel
DILLENBECK, DOUGLAS D., College Entrance Examination Board
DION, ROBERT, California Test Bureau
DOBBIN, JOHN E., Educational Testing Service
DONLON, THOMAS F., Educational Testing Service
DOPPELT, JEROME E., The Psychological Corporation
DOWNES, MARGARET C., New York State Department of Education
DRAGOSITZ, ANNA, Educational Testing Service
DREW, MARY L., Educational Testing Service
DREWS, ELIZABETH M., Michigan State University
DRY, RAYMOND J., Life Insurance Agency Management Association
DUBNICK, LESTER, Huntington (New York) Public Schools
DUFFORD, JOHN R., JR., The Pingry School
DUNCANSON, JAMES, Educational Testing Service
DUNN, FRANCES E., Brown University
DURHAM, JOYCE, The King's College
DUROST, WALTER N., Test Service and Advisement Center
DUTTON, EUGENE, Rhode Island College
DYER, HENRY S., Educational Testing Service

EAGLE, NORMAN, Fort Lee (New Jersey) Public Schools
EBEL, ROBERT L., Michigan State University
EDWARDS, WINIFRED E., Irvington (New Jersey) High School
EGAN, JEROME, St. Francis College
EICHLER, HERBERT, Teachers College, Columbia University
EISENBERG, I., New York City Department of Personnel
ELAINE, SISTER M., S.S.N.D., Our Lady Queen of Heaven School (Louisiana)
ELLIOTT, GODFREY, McGraw-Hill Book Company
ELLIOTT, MERLE H., Oakland (California) Public Schools
ENGLEHART, MAX D., Chicago (Illinois) City Junior College
ENGLEHART, MRS. MAX D., South Shore High School, Chicago, Illinois
EPSTEIN, BERTRAM, The City College of New York
FAN, CHUNG T., International Business Machines Corporation
FARR, S. DAVID, State University of New York at Buffalo
FEIFER, IRWIN, Institute for Crippled and Disabled (New York City)
FELDMANN, SHIRLEY, The City College of New York
FELDT, LEONARD S., University of Iowa
FENDRICK, PAUL, Millburn, New Jersey
FENOLLOSA, GEORGE M., Houghton Mifflin Company
FENSTERMACHER, GUY M., Educational Testing Service
FERRIS, ANNE H., Educational Testing Service
FERRIS, FREDERICK L., JR., Princeton University
FICCA, CHARLES, Monmouth College
FIELDER, EARL R., Educational Testing Service
FIFER, GORDON, Hunter College
FINDLEY, WARREN G., University of Georgia
FINEGAN, OWEN T., Gannon College
FINK, DAVID R., JR., University of Maine
FINNERTY, M., New York University
FITZGERALD, JOHN F. M., Hillcrest School, Brookline, Massachusetts
FLANAGAN, JOHN C., American Institute for Research
FLAUGHER, RONALD L., Educational Testing Service
FLEENER, DONALD E., Indiana Central College
FLEISCH, SYLVIA, Boston University
FLESCH, MARY H., Educational Testing Service
FLEMMING, EDWIN G., New York City
FLYNN, JOHN T., University of Connecticut
FONSEA, JOHN A., Broncksland Junior High School, Bronx, New York
FORLANO, ANN MARIE, New York City Public Schools
FORLANO, GEORGE, New York City Board of Education

FORMICA, LOUIS A., Norwalk (Connecticut) Public Schools
FORRESTER, GERTRUDE, H. W. Wilson Company
FOXE, ESTHER K., University of Maryland
FRANCIS, BROTHER COSMAS, F.S.C., Brothers of the Christian Schools, Narragansett, Rhode Island
FREDERICKSEN, JOHN, Educational Testing Service
FREDERICKSEN, NORMAN, Educational Testing Service
FREEMAN, PAUL M., Princeton, New Jersey
FRENCH, BENJAMIN J., New York State Department of Civil Service
FRENCH, JOHN W., Educational Testing Service
FRENCH, ROBERT L., Science Research Associates
FRICKE, BENNO G., University of Michigan
FRIEDERMAN, ROBERT, New Jersey State Department of Civil Service
FRIEDMAN, SIDNEY, Bureau of Naval Personnel, Washington, D. C.
FRUCHTER, BENJAMIN, The University of Texas
FUCHS, EDMUND F., U. S. Army Personnel Research Office
FULTON, RENEE J., New York City Board of Education
GALLAGHER, HENRIETTA L., Educational Testing Service
GANNON, FRED, Educational Testing Service
GARDNER, ERIC F., Syracuse University
GASIOROWSKI, MARY, City University of New York
GEE, HELEN H., National Institute of Health
GEORGIA, SISTER M., Rosary Hill College, Buffalo, New York
GERJUOY, HERBERT, Educational Testing Service
GIBLETTE, JOHN F., University of Maryland
GILLETTE, ANNETTE L., Board of Education, Hartford, Connecticut
GILLEN, W. KING, Harvard University
GLASIER, CHARLES A., New York State Department of Education
GLASS, DAVID C., Russel Sage Foundation
GLESER, GOLDINE C., University of Cincinnati
GLICKMAN, ALBERT S., United States Department of Agriculture
GOBETZ, WALLACE, New York University
GODSHALK, FRED I., Educational Testing Service
GOLDMAN, LEO, Brooklyn College
GOLDSTEIN, LEO S., New York Medical College
GOLDSTEIN, WILLIAM, Central School District No. 4, Plainview, New York
GORDON, LEONARD V., United States Army Personnel Research Office
GOREN, ARNOLD L., New York University
GOSLIN, DAVID A., Russell Sage Foundation
GOTKIN, ELIZABETH, Teachers College, Columbia University

GOTKIN, LASSAR G., New York Medical College
GOWER, JENETTE, St. Mary's School, Peekskill, New York
GOWING, JAMES D., Tabor Academy, Marion, Massachusetts
GRAFF, FRANKLYN A., Westport (Connecticut) Public Schools
GRANT, DONALD L., American Telephone and Telegraph Company
GRAVES, WALTER A., NEA Journal
GRAY, LYLE BLAINE, Diagnostic and Remedial Center, Baltimore, Maryland
GREEN, DOROTHY, United States Civil Service Commission
GREENWALD, ANTHONY G., Educational Testing Service
GREER, HARRY H., JR., Naval Reserve Officers Training Corps
GUERRIERO, MICHAEL A., The City College of New York
GULLIKSEN, HAROLD, Educational Testing Service
GUMMERE, JOHN F., William Penn Charter School, Philadelphia, Pennsylvania
GUTHRIE, GEORGE M., Pennsylvania State University
HADLEY, EVERETT E., West Hartford (Connecticut) Public School
HAGGERTY, HELEN R., United States Army Personnel Research Office
HAGMAN, ELMER R., Greenwich (Connecticut) Public Schools
HALL, ROBERT G., Manter Hall School, Cambridge, Massachusetts
HALL, ROY M., University of Delaware
HAMBURGER, MARTIN, New York University
HAMEL, LESTER S., Pennsylvania State University
HARDESTY, ANNE S., Biometrics Research
HAROOTUNIAN, BERJ, University of Delaware
HARTSHORNE, NATHANIEL H., Educational Testing Service
HARVEY, PHILIP R., Educational Testing Service
HASTINGS, J. THOMAS, University of Illinois
HAVEN, ELIZABETH, Educational Testing Service
HAWES, GENE, New York City
HAYES, MARY E., United States Office of Education
HAYWARD, PRISCILLA, Educational Testing Service
HAZZARD, MARY E., New York University
HEATH, C. NEWTON, Lincoln-Sudbury (Massachusetts) Regional School District
HEIDGERD, LLOYD H., Educational Testing Service
HEIL, LOUIS, Brooklyn College
HEISER, RUTH BISHOP, Glendale, Ohio
HELMICK, JOHN S., Educational Testing Service
HEMPHILL, JOHN K., Educational Testing Service
HENRY, SALLYANN, The Psychological Corporation

HERMAN, DAVID, The Psychological Corporation
HERRICK, C. JAMES., Hartwick College
HESLIN, PHYLLIS, National League for Nursing
HIERONYMUS, ALBERT N., University of Iowa
HILLS, JOHN R., University System of Georgia
HILTON, THOMAS L., Educational Testing Service
HINDSMAN, EDWIN, Indiana University
HITCHCOCK, ARTHUR A., American Personnel and Guidance Association
HOCHMAN, SIDNEY, Queens College
HOFFMAN, E. LEE, Tulane University
HOFFMAN, HELMUT, Educational Testing Service
HOFSTEE, WILLEM K. B., Educational Testing Service
HOLLENBECK, GEORGE, International Business Machines Corporation
HOLLIS, ESTHER R., The Psychological Corporation
HOLLISTER, JOHN S., Educational Testing Service
HONAKER, LINTON R., Tuscarawas County (Ohio) Schools
HOPMANN, ROBERT P., The Lutheran Church – Missouri Synod
HOPKINS, FLORENCE M., Educational Testing Service
HORN, DOROTHY M., Teachers College, Columbia University
HOROWITZ, GEORGE F., Columbia University
HOROWITZ, LEOLA S., Adelphi University
HOROWITZ, MILTON W., Queens College
HOWARD, ALLEN H., University of Illinois
HOWELL, JOHN, Educational Testing Service
HUBBARD, JOHN P., National Board of Medical Examiners
HUGHES, JOHN L., International Business Machines Corporation
HUMPHRY, BETTY J., Educational Testing Service
HUNT, G. HALSEY, Educational Council for Foreign Medical Graduates
HUTCHINSON, MARY G., Wilby High School, Waterbury, Connecticut
HUYSER, ROBERT J., Educational Testing Service
HUYSER, SARAH L., Educational Testing Service
IMPELLIZZERI, IRENE H., Brooklyn College
IRBY, ALICE J., Educational Testing Service
IRVINE, PAUL, JR., Pennsylvania Department of Public Instruction
JACKSON, DOUGLAS N., Stanford University
JACOBS, ROBERT, Southern Illinois University
JACOBS, PAUL, Educational Testing Service
JAMES, GRACE ROBBINS, University of North Carolina
JANEBA, HUGO B., Rutherford (New Jersey) Senior High School
JARECKE, WALTER H., West Virginia University

149

JASPEN, NATHAN, New York University
JOHNSON, A. PEMBERTON, Newark College of Engineering
JOHNSON, DAVID, Knowlton School, Bronx, New York
JOHNSON, RICHARD, Rutgers, The State University
JONES, LYLE V., University of North Carolina
JUOLA, ARVO E., Michigan State University
KABACK, GOLDIE R., The City College of New York
KAMMAN, JAMES F., University of Illinois
KARAS, SHAWKY F., Educational Testing Service
KARL, MADELINE, New York City Board of Education
KATHLEEN, SISTER MARY, College of St. Elizabeth
KATZ, MARTIN R., Educational Testing Service
KATZELL, MRS. RAYMOND A., National League for Nursing
KAUFFMAN, ELEONORA J., The University of Chicago
KELEHER, REVEREND GREGORY, St. Anselm's College
KELLEY, PAUL R., National Board of Medical Examiners
KELLEY, H. PAUL, University of Texas
KELLEY, MRS. H. PAUL, Austin, Texas
KELLY, E. LOWELL, University of Michigan
KENDALL, LORNE M., Educational Testing Service
KENDALL, W. E., The Psychological Corporation
KENNEDY, S. M., Texas Technological College
KENNEY, HELEN I., Harvard University
KENT, CLARENCE E., Virginia State Department of Education
KERSTING, ETHEL F., Educational Testing Service
KESSMAN, MAURICE, Rochester Institute of Technology
KIEFFER, JOHN E., Harker Preparatory School, Potomac, Maryland
KIRKPATRICK, FORREST H., Wheeling Steel Corporation
KLINE, WILLIAM E., Baltimore County (Maryland) Public Schools
KLING, FREDERICK R., Educational Testing Service
KOCH, JOHN C., JR., Madison (New Jersey) High School
KOGAN, LEONARD S., Brooklyn College
KOGAN, NATHAN, Educational Testing Service
KOOB, REVEREND C. ALBERT, National Catholic Educational Association
KRIGSMAN, RUBEN, New York City Community College
KRUG, ROBERT E., American Institute for Research
KUJAWSKI, CARL J., The Atlantic Refining Company
KUNOFSKY, NORMA, New York State Department of Civil Service
KUNOFSKY, SOLOMON, New York State Department of Health
KURFMAN, DANA G., Educational Testing Service

KURLAND, NORMAN D., New York State Department of Education
KVARACEUS, WILLIAM C., Tufts University
LAGRONE, HERBERT F., American Association of Colleges for Teacher Education
LAIDLAW, WILLIAM J., Hunter College
LAMBERT, JANE E., Educational Testing Service
LANDY, EDWARD, Newton (Massachusetts) Public Schools
LANE, HUGH W., The University of Chicago
LANE, WILLIAM S., Vashon Island High School, Burton, Washington
LANGMUIR, C. R., The Psychological Corporation
LANNHOLM, GERALD V., Educational Testing Service
LATHROP, ROBERT L., Pennsylvania State University
LAUDAN, BONNIE S., Educational Testing Service
LAVINE, MARIAN, Jericho (New York) High School
LEES, DIANA M., Educational Testing Service
LEHMKUHL, CARLTON B., Boston College
LEIDNER, BURTON R., International Business Machines Corporation
LEIGHTON, P. L., East Stroudsburg State College, Pennsylvania
LENNON, ROGER T., Harcourt, Brace and World, Inc.
LEVERETT, HOLLIS M., Hollis M. Leverett and Associates, Inc.
LEVINE, ABRAHAM S., Office of Naval Research, Washington, D. C.
LEVINE, HAROLD G., New York State Department of Education
LEVINE, MILTON, National Science Foundation
LIBERMAN, SAMUEL S., Columbia University
LILLY, ROY S., Educational Testing Service
LINDBERG, LUCILE, Queens College
LINDEMAN, RICHARD H., Teachers College, Columbia University
LINDQUIST, E. F., State University of Iowa
LINDVALL, C. M., University of Pittsburgh
LINK, FRANCES R., Cheltenham Township (Pennsylvania) Schools
LINTON, LINDA, New York City Public Schools
LOHNES, PAUL R., State University of New York at Buffalo
LONG, LOUIS, The City College of New York
LONG, WILLIAM F., United States Air Force
LOREE, M. RAY, University of Alabama
LORET, PETER G., Educational Testing Service
LORETAN, JOSEPH O., New York City Board of Education
LOTTES, JOHN J., State University of New York at Geneseo
LOWERY, ZEB A., Rutherford County (North Carolina) Board of Education
LUCAS, DIANA D., Educational Testing Service

LUCAS, NORMA LEE, Clayton (Missouri) School District
LUTZ, ORPHA M., Montclair State College, New Jersey
LYMAN, HOWARD B., University of Cincinnati
LYNCH, ELEANOR, National League for Nursing
LYONS, WILLIAM A., New York State Department of Education
MAC BAIN, ROBERT T., The Torrington Company
MADAUS, GEORGE F., State College of Worcester, Massachusetts
MADDI, SALVATORE R., Educational Testing Service
MADDOX, CLIFFORD R., Cedarville College
MAIER, MILTON H., Educational Testing Service
MALCOLM, DONALD J., Educational Testing Service
MALONEY, DANIEL J., New York State Department of Education
MARKOWICZ, REVEREND WALTER A., Sacred Heart Seminary, Detroit, Michigan
MARRON, JOSEPH E., United States Military Academy
MARSH, JAMES V., Educational Testing Service
MARX, GEORGE L., University of Maryland
MASON, ANDREW T., Maryland State Department of Education
MASSEY, WILL J., University of Maryland
MATHEWSON, ROBERT H., The City University of New York
MAYER, MARTIN, New York City
MAYFIELD, EUGENE C., Life Insurance Agency Management Association
MC CALL, W. C., University of South Carolina
MC CANN, FORBES E., McCann Associates, Philadelphia
MC CARTHY, DOROTHEA, Fordham University
MC CONNELL, JOHN C., Windward School, White Plains, New York
MC CORD, RICHARD B., Philadelphia Personnel Department
MC CULLERS, WAYNE M., New York City Community College
MC DANIEL, SARAH W., Hofstra University
MC DILL, THOMAS H., The Westminister Schools, Atlanta, Georgia
MC GUIRE, CHRISTINE, University of Illinois
MC GUIRE, JOSEPH, New Jersey State Department of Civil Service
MC INTIRE, PAUL H., Boston University
MC KEE, MICHAEL G., United States Government
MC KENNA, MAE R., Crosby High School, Waterbury, Connecticut
MC KENZIE, FRANCIS W., Board of Education, Darien, Connecticut
MC KEON, JAMES J., Educational Testing Service
MC LAUGHLIN, KENNETH F., United States Office of Education
MC LEAN, LESLIE DAVID, University of Wisconsin
MC MANUS, LEO F., JR., Cleaver Company Executive Institute

MCMULLIN, THOMAS E., University of Pennsylvania
MC NULTY, THEODORE F., Educational Testing Service
MC PEEK, BUCKNAM, Massachusetts General Hospital
MC PEEK, W. MILES, Harvard University
MC QUITTY, JOHN V., University of Florida
MC TARNAGHAN, ROY E., State University of New York at Geneseo
MEDLEY, DONALD M., The City University of New York
MEDRANO, LOURDES M., New York University
MEHDI, BAQER, National Institute of Education, New Delhi, India
MELTON, RICHARD S., Educational Testing Service
MELVILLE, S. DONALD, Educational Testing Service
MERNYK, CHARLOTTE LEVY, Brooklyn, New York
MERRITT, ROBERT T., College Entrance Examination Board
MERWIN, JACK C., University of Minnesota
MESSICK, SAMUEL J., Educational Testing Service
MICHELI, GENE, United States Naval Training Device Center
MIDDENDORF, LORNA, Roosevelt Junior High School, Westfield, New Jersey
MILES, NELLE H., Falls Church (Virginia) High School
MILLER, BEN F., III, The Psychological Corporation
MILLER, DOROTHY F., Clayton (Missouri) School District
MILLER, HOWARD G., North Carolina State College
MILLER, E. JOYCE, New York University
MILLER, MARIAN B., Delaware State Department of Public Instruction
MILLER, PAUL VANR., JR., Educational Testing Service
MILLMAN, JASON, Cornell University
MILLS, DONALD F., Educational Testing Service
MIRKIN, LOUISE, Educational Testing Service
MITCHELL, BLYTHE C., Harcourt, Brace and World, Inc.
MITZEL, HAROLD E., Pennsylvania State University
MOHNACS, ANNA, National Catholic Educational Association
MOHNACS, MARY, National Catholic Educational Association
MOLLENKOPF, WILLIAM G., Procter and Gamble Company
MOORE, MAXINE R., Educational Testing Service
MORGAN, HENRY H., The Psychological Corporation
MORIARTY, DORIS, Educational Testing Service
MORRISON, ALEXANDER W., Polytechnic Institute of Brooklyn
MOSELY, RUSSELL, Wisconsin State Department of Public Instruction
MUKHERJEE, MRS. G., Windsor Mountain School, Lenox, Massachusetts
MULRY, JUNE, Indiana University
MURRAY, VIRGINIA E., Educational Testing Service

MYERS, CHARLES T., Educational Testing Service
MYERS, ISABEL BRIGGS, Swarthmore, Pennsylvania
NATHAN, CYNTHIA R., United States Office of Health, Education and Welfare
NELSON, GID E., University of South Florida
NELSON, H. ROBERT, Highland Park (New Jersey) High School
NEVIN, MARGARET H., Educational Testing Service
NOISEUX, ETHEL R., New York State Department of Civil Service
NOLAN, DAVID M., Educational Testing Service
NOLL, VICTOR H., Harcourt, Brace and World, Inc.
NORA, SISTER MARY, S.S.N.D., National Catholic Educational Association
NORTH, ROBERT D., Educational Records Bureau
NORTON, ELIZABETH, Teachers College, Columbia University
NORTON, MARGARET, Hopewell, New Jersey
NOSOW, SIGMUND, Michigan State University
NULTY, FRANCIS X., Educational Testing Service
OHNMACHT, FRED W., University of Maine
OILL, B. C., New York City Department of Personnel
O'KEEFE, JOHN J., Science Research Associates
OPPENHEIM, DON B., Teachers College, Columbia University
ORAHOOD, ELIZABETH, Kansas City (Missouri) Public Schools
OSTRAM, ELIZABETH, New York State Department of Civil Service
ORLEANS, JOSEPH B., George Washington High School, New York City
ORR, DAVID B., American Institute for Research
OSCARSON, DONALD, The Taft School, Watertown, Connecticut
OSGOOD, STANLEY W., Houghton Mifflin Company
OTIS, C. ROBERT, California Test Bureau, Fulton, New York
OTTOBRE, FRANCES M., Educational Testing Service
OWENS, ROBERT G., State University of New York at Buffalo
PACKARD, ALBERT G., Baltimore City (Maryland) Public Schools
PAGE, ELLIS B., University of Connecticut
PALLRAND, GEORGE J., Princeton University
PALMER, ORMOND E., Michigan State University
PALMER, ORVILLE B., Educational Testing Service
PALUBINSKAS, ALICE L., Tufts University
PAPPAS, ANGELINE J., Horace Greeley High School, Chappaqua, New York
PATTERSON, SUSAN, New York University
PATRICE, SISTER M., O.S.F., Cardinal Stritch College
PAYNE, DAVID A., Syracuse University
PELIKAN, PHYLLIS K., New York University

PERRY, WILLIAM D., University of North Carolina
PETERSON, DONALD A., Life Insurance Agency Management Association
PETERSON, RICHARD E., Educational Testing Service
PFEIFFER, ANN, Baltimore City (Maryland) Public Schools
PIERSON, ELLERY M., Educational Testing Service
PITCHER, BARBARA, Educational Testing Service
POLLACK, NORMAN C., New York State Civil Service Department
POOLE, RICHARD L., Syracuse University
POOLER, MARY H., Erie (Pennsylvania) School District
PRESTON, BRAXTON, Educational Testing Service
PRUZEK, FRANK, University of Wisconsin
PRUZEK, ROBERT, University of Wisconsin
PURCELL, WILLIAM D., Summit (New Jersey) Public Schools
PURDY, ROBERT D., Syracuse City (New York) School District
PUTZIG, ALBERT, Philadelphia Personnel Department
QUINN, JOHN S., JR., Harcourt, Brace and World, Inc.
RA, JUNG BAY, Harcourt, Brace and World, Inc.
RABINOWITZ, WILLIAM, The City University of New York
RANDOLPH, LAWRENCE, Tenafly (New Jersey) High School
RAPHAEL, BROTHER ALOYSIUS, F.S.C., Bishop Loughlin Memorial High School, Brooklyn
RAPPARLIE, JOHN H., The Owens-Illinois Glass Company
READ, THOMAS, Hampton Roads Academy, Newport News, Virginia
REEBER, MARY K., Educational Testing Service
REED, BERNARD A., Trenton State College
REED, REVEREND LORENZO K., S.J., Jesuit Educational Association
REELING, GLENN E., Montclair (New Jersey) Public Schools
REID, CATHERINE F., Hunter College
REID, JOHN W., Indiana State College, Pennsylvania
REILLY, JAMES J., St. John's University
REISS, JEAN F., Educational Testing Service
REUTER, WILLIAM H., Educational Testing Service
REYNOLDS, HARLAN J., International Business Machines Corporation
RHODES, DORIS L., Educational Testing Service
RHUM, GORDON J., State College of Iowa
RICHARDINE, SISTER MARY, B.V.M., National Catholic Educational Association
RICHARDS, JAMES M., Educational Testing Service
RICKS, JAMES H., JR., The Psychological Corporation
ROBERTS, G. TRACEY, Pennsylvania State Civil Service Commission

ROBBINS, IRVING, Queens College
ROBINSON, DONALD W., Phi Delta Kappa
ROCK, ROBERT T., Harcourt, Brace and World, Inc.
ROHRBAUGH, FRANCES G., Educational Testing Service
ROHRBAUGH, JOSEPH W., Pennsbury Senior High School, Yardley, Pennsylvania
ROMBERG, THOMAS, National Longitudinal Study of Mathematical Ability
ROSEBOROUGH, HOWARD, McGill University
ROSEN, JULIUS, New York City Public Schools
ROSNER, BENJAMIN, Brooklyn College
ROSS, WESLEY F., University of Kentucky
ROSSER, DONALD S., New Jersey Education Association
ROWLAND, WILMINA, The United Presbyterian Church, Board of Christian Education
SACHS, LORRAINE P., National League for Nursing
SANAZARO, PAUL J., Association of American Medical Colleges
SANBORN, MARSHALL P., University of Wisconsin
SANFORD, RUTH C., West Hempstead (New York) Junior-Senior High School
SASAJIMA, MASU, Educational Testing Service
SASLOW, MAX S., New York City Department of Personnel
SCHEIDER, ROSE M., Educational Testing Service
SCHLEKAT, GEORGE A., Educational Testing Service
SCHNEIDER, HARRIET L., National League for Nursing
SCHNITZEN, JOSEPH P., University of Houston
SCHRADER, WILLIAM B., Educational Testing Service
SCHULTZ, CHARLES B., Educational Testing Service
SCHULTZ, DOUGLAS, Applied Psychological Services
SCHULZ, DELPHIN L., The Lutheran Church-Missouri Synod
SCHUMACHER, CHARLES F., National Board of Medical Examiners
SCHWARTZMAN, ALEX E., McGill University
SCOFIELD, LEONARD, Dean Junior College
SCOTT, C. WINFIELD, Rutgers, The State University
SCOTT, MARY HUGHIE, National Education Association
SCRIBNER, PETER C., Harcourt, Brace and World, Inc.
SEASHORE, HAROLD, The Psychological Corporation
SEIBEL, DEAN W., Educational Testing Service
SENFT, ALBERT, Shaker Heights (Ohio) High School
SERAFINO, ROBERT P., Educational Testing Service
SERLING, ALBERT M., Educational Testing Service

SETZER, CHARLES J., New York City Department of Personnel
SFORZA, RICHARD F., United States Military Academy
SHAPIRO, BERNARD J., Harvard University
SHARP, CATHERINE G., Educational Testing Service
SHAYCOFT, MARION F., American Institute for Research
SHEA, JAMES E., New York State Department of Civil Service
SHEEHE, MARCIA, University of Rochester
SHERMAN, EDGAR M., Irvington (New Jersey) High School
SHIELDS, MARY, National League for Nursing
SHIELDS, O. L., Jefferson County (Kentucky) Education Center
SHIMBERG, BENJAMIN, Educational Testing Service
SIEGEL, ARTHUR, Applied Psychological Services
SIEGEL, LAURENCE, Miami University
SILVEY, HERBERT M., State College of Iowa
SIMANDLE, SIDNEY, Kentucky State Department of Education
SJOBERG, LENNART, Educational Testing Service
SKAGER, RODNEY W., Educational Testing Service
SLOAT, CHESTER H., Pennsylvania Military College
SMITH, ALBERT C., United States Marine Corps
SMITH, ALEXANDER F., Southern Connecticut State College
SMITH, ALLAN B., Rhode Island College
SMITH, ANN Z., Educational Testing Service
SMITH, JOHN P., Educational Testing Service
SMITH, MARSHALL P., Trenton State College
SMITH, MARSHALL S., Harvard University
SMITH, ROBERT E., Educational Testing Service
SMITH, VIRGINIA A., New York University
SNODGRASS, ROBERT, Purdue University
SOLOMON, ROBERT J., Educational Testing Service
SOMMER, JOHN, Houghton Mifflin Company
SOUTHER, MARY T., Tower Hill School, Wilmington, Delaware
SOUTHWORTH, J. ALFRED, University of Massachusetts
SPAIN, CLARENCE J., Schenectady (New York) Public Schools
SPEER, GEORGE S., Illinois Institute of Technology
SPENCE, JAMES R., State University of New York at Albany
SPENCER, RICHARD E., Pennsylvania State University
SPITZER, ROBERT L., Biometrics Research
SPRAGUE, ARTHUR R., Hunter College
SPREMULLI, ESTELLE E., Educational Testing Service

SQUIRES, JO ANNE S., Hampton Roads Speech and Hearing Center, Newport News, Virginia
STAHL, SUZANNE, Educational Testing Service
STAKE, ROBERT E., University of Illinois
STAMM, MARTIN L., Educational Testing Service
STANTON, E. E., JR., University of South Florida
STARK, ALICE, New York City Public Schools
STATLER, CHARLES R., State University of Iowa
STEINMAN, ARTHUR MILES, Trenton State College
STERN, GEORGE G., Syracuse University
STERN, JACK I., New York City Department of Personnel
STEWART, BLAIR, Associated Colleges of the Midwest
STEWART, CLIFFORD T., University of South Florida
STEWART, E. ELIZABETH, Educational Testing Service
STEWART, NAOMI S., Cranford, New Jersey
STICE, GLEN, New York City
STICKELL, DAVID W., Educational Testing Service
STILES, GRACE ELLEN, University of Rhode Island
STOKER, HOWARD W., Florida State University
STONE, PAUL T., Huntingdon College
STOUGHTON, ROBERT W., Connecticut State Department of Education
STREICHER, SAMUEL, New York City Board of Education
STRIBULA, MICHAEL, Educational Testing Service
STRICKER, LAWRENCE J., Educational Testing Service
STUARDI, MONSIGNOR J. EDWIN, Diocese of Mobile-Birmingham, Alabama
STUDDIFORD, WALTER B., Princeton University
STUNKARD, CLAYTON L., University of Maryland
SUPER, DONALD E., Teachers College, Columbia University
SUSSMAN, LEAH, Newark (New Jersey) School System
SUTTON, JOSEPH T., Stetson University
SUVAK, ALBERT, Montana State College
SWAN, BEVERLY B., Florida State Department of Education
SWANSON, EDWARD O., University of Minnesota
SWEENEY, EDWARD, Life Insurance Agency Management Association
SWINEFORD, FRANCES, Educational Testing Service
TAYLOR, BRUCE L., Educational Testing Service
TAYLOR, MARVIN, Queens College
TAYLOR, SAMUEL J., New York State Civil Service Department
TERRAL, JOSEPH E., Educational Testing Service
THOMAS, WILLIAM F., University of Wisconsin

THOMPSON, RAYMOND E., Educational Testing Service
THORNDIKE, ROBERT L., Teachers College, Columbia University
THORNE, MARGARET A., Educational Testing Service
TRAUB, ROSS E., Educational Testing Service
TRAXLER, ARTHUR E., Educational Records Bureau
TRAXLER, MRS. ARTHUR, New York City
TREXLER, LAURA M., Northern Valley Regional High School, Demarest, New Jersey
TRIGGS, FRANCES, Committee on Diagnostic Reading Tests, Inc.
TRISMEN, DONALD A., Educational Testing Service
TUCKER, DONALD K., Northeastern University
TUCKER, LEDYARD R, University of Illinois
TULLY, G. EMERSON, Florida State University
TURNBULL, WILLIAM W., Educational Testing Service
TWYFORD, LORAN C., New York State Department of Education
TYLER, MATILDA, Simpson School, Wallingford, Connecticut
UMANS, SHELLEY, New York City Board of Education
UPSHALL, CHARLES C., Eastman Kodak Company
URQUHART, HELEN L., Brown University
VALENTINE, JOHN A., College Entrance Examination Board
VALLEY, JOHN R., Educational Testing Service
VAN AUSDALL, LEIGH, Science Research Associates
VAN HORN, MARY JANE, Baltimore County (Maryland) Schools
VEZINA, BEVERLY, Educational Testing Service
VISCEGLIA, JOHN A., Camden City (New Jersey) Schools
VON FANGE, ERICH A., Concordia College
WADELL, BLANDENA C., Harcourt, Brace and World, Inc.
WAGNER, E. PAUL, Bloomsburg State College, Pennsylvania
WAGNER, WILLIAM, West Virginia University
WAHLGREN, HARDY L., State University of New York at Geneseo
WALDEN, CARRIE H., New York University
WALKER, HOWARD, Mt. Vernon (New York) Public Schools
WALKER, ROBERT N., Personnel Press, Princeton, New Jersey
WALLACE, WIMBURN L., The Psychological Corporation
WALLACH, PHILIP C., Wallach Associates, Inc.
WALTHER, REGIS, United States Department of State
WALTHEW, JOHN K., Educational Testing Service
WALTON, WESLEY W., Educational Testing Service
WALTZMAN, HAL, New York State Department of Civil Service
WANTMAN, MOREY ., Educational Testing Service

WARD, ANNIE W., Volusia County (Florida) Board of Public Instruction
WASSERMANN, LORRAINE, Educational Testing Service
WATKINS, RICHARD W., Educational Testing Service
WATSON, WALTER S., The Cooper Union
WEBB, SAM C., Emory University
WEEKS, HAROLD L., San Francisco (California) Schools
WEINER, MAX, Brooklyn College
WEISS, JOSEPH, Polytechnic Institute of Brooklyn
WESMAN, ALEXANDER G., The Psychological Corporation
WESMAN, MRS. ALEXANDER G., New York City
WHITE, HERBERT L., New York City Department of Personnel
WHITELAW, JOHN B., United States Department of Health, Education and Welfare
WHITLA, DEAN K., Harvard University
WHITNEY, ALFRED G., Life Insurance Agency Management Association
WIENER, SOLOMON, New York City Department of Personnel
WILEY, ISABEL C., Pennsbury High School, Yardley, Pennsylvania
WILKE, MARGUERITE, New York University
WILKE, WALTER H., New York University
WILLARD, RICHARD W., Massachusetts Institute of Technology
WILLAUER, PETER O., Groton (Massachusetts) School
WILLEMIN, LOUIS P., United States Army Personnel
WILLIAMS, HENRY A., JR., Pensacola (Florida) Junior College
WILLIAMS, ROGER K., Morgan State College, Maryland
WILLEY, CLARENCE F., Norwich University
WILSON, MARILYN, Educational Testing Service
WINGO, ALFRED L., Virginia State Department of Education
WINIEWICZ, CASIMER S., United States Naval Examining Center
WINTERBOTTOM, JOHN A., Educational Testing Service
WISE, HAROLD L., Western Reserve University
WOEHLKE, ARNOLD B., International Business Machines Corporation
WOLF, RICHARD, University of Chicago
WOOD, BEN D., Columbia University
WOOLLATT, LORNE H., New York State Department of Education
WRIGHT, WILBUR H., State University of New York at Geneseo
WRIGHTSTONE, J. WAYNE, New York City Board of Education
YABLONSKY, FRANK D., New York City
YABLONSKY, SHIRLEY, New York University
YOXALL, GEORGE J., Inland Steel Company
ZACCARIA, LUCY C., University of Illinois

ZALKIND, SHELDON S., The City College of New York
ZARRO, PASQUALE J., Philadelphia Personnel Department
ZEFF, LEON H., Board of Education, Darien, Connecticut
ZIMILES, HERBERT, Bank Street College
ZOLA, EUGENE J., Niskayuna High School, Schenectady, New York
ZOOK, DONOVAN Q., Board of Examiners for the Foreign Service
ZUCKERMAN, HAROLD, New York City Board of Education